DOCUMENT RESUME

ED 302 707 CE 051 739

TITLE Probation Officer Validation Report and Technical Adequacy Report. Standards and Training for Corrections Program.
INSTITUTION California State Board of Corrections, Sacramento.
PUB DATE 87
NOTE 131p.; For related documents, see CE 051 717-725, 729-735, and 740-744.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Adult Education; Correctional Rehabilitation; Criterion Referenced Tests; Employment Practices; Entry Workers; Evaluation Criteria; *Occupational Tests; *Personnel Selection; Postsecondary Education; Predictive Validity; *Probation Officers; Research Methodology; State Standards; *Statewide Planning; Test Construction; Testing; *Test Validity
IDENTIFIERS *California

ABSTRACT

In response to a mandate from the California State Legislature, the California Board of Corrections conducted a three-phase research study in preparation for the revision and standardization of the current state standards for hiring and training corrections personnel. The primary focus of the second phase of the project was on the development and validation of entry-level selection procedures for these positions. This set of technical reports deals with the process of developing and validating the criteria on which the selection procedures developed were based. The validation report covers the following topics: the project goals and the research strategies used to achieve them; the principal results of the job analysis; the predictor and criterion development processes; data collection procedures used; the criterion, predictor, and validity and fairness analyses used in developing the selection battery; and recommendations pertaining to administering and scoring the test. The technical adequacy report discusses the validity of the testing program in view of the fact that, although the original project plans called for testing a minimum of 500 probation officers, the final validation sample ranged from 284 to 322 officers depending on the analysis being conducted. (MN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302707

PROBATION OFFICER VALIDATION REPORT AND TECHNICAL ADEQUACY REPORT

STANDARDS AND TRAINING FOR CORRECTIONS PROGRAM

CE051739

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C.N. Page

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

STANDARDS AND TRAINING
FOR CORRECTIONS PROGRAM

STC

BOARD OF CORRECTIONS

VALIDATION REPORT

PROBATION OFFICER

Copyright 1987 California Board of Corrections

In preparing the Probation Officer Examination for the California Board of Corrections, Personnel Decisions, Inc. has taken into account the following documents concerning selection testing:

- The American Psychological Association, Standards for Educational and Psychological Tests, (1974).
- The APA Division of Industrial-Organizational Psychology, Principles for the Validation and Use of Personnel Selection Procedures, 2nd edition, (1980).
- The Equal Employment Opportunity Commission, Uniform Guidelines on Employee Selection Procedures, (1978).

CONTENTS

Page

List of Tables ii
List of Figures iv

Chapter 1 INTRODUCTION

Project Background..... 1
Participating Departments..... 2
Research Strategy and Rationale..... 2
Report Overview..... 4

Chapter 2 SUMMARY OF THE JOB ANALYSIS 5

Chapter 3 PREDICTOR DEVELOPMENT

Overview and Rationale..... 7
Linking KSAs to Job Tasks..... 9
Grouping Task Items into Task Categories.....13
Task Category to Worker Characteristic Ratings.....13
Setting Linkage Criteria.....20
Linking Tests to Worker Characteristics.....20
Linking Tests to Task Categories.....31
Predictor Development.....36
Test construction.....36
Pretesting the preliminary tests.....38
Item analysis and revision.....38
Special revision methods.....41
Final Test Battery.....42
Noncognitive Tests.....43
Sample Items Booklet.....44

Chapter 4 CRITERION DEVELOPMENT

Collecting Critical Incidents.....45
Revising and Retranslating the Critical Incidents.....46
Developing Behavioral Summary Scales.....51
Other Ratings Scales.....51
Worker characteristic ratings.....51
Overall effectiveness.....53

Chapter 5 DATA COLLECTION

Predictor Data Collection.....54
Criterion Data Collection.....55
Final Validation Sample.....56
Missing name.....56

CONTENTS (Chapter 5 continued)

Page

Missing ratings....56
Low tenure....58
Inaccurate ratings....58
Demographics of Final Validation Sample....58
Technical Adequacy of Sample....62

Chapter 6 ANALYSES AND RESULTS

Criterion Analyses....64
Data collection and cleaning....64
Developing an overall criterion....66
Search for additional criteria....69
Criterion descriptive statistics....72
Predictor Analyses....76
Data collection and cleaning....76
Scale revisions and final psychometrics....76
Descriptive statistics....77
Validity and Fairness Analyses....77
Adverse impact and test fairness....84
Validity and fairness of the individual tests....85
Developing a Final Selection Battery....88
Differential predictability of specific criteria....89

Chapter 7 DISCUSSION AND RECOMMENDATIONS

Exam Format and Content....93
Using the Probation Officer Examination....96
Exam Security....96
Scoring and Data Collection....97
Score Interpretation and Cutoff Scores....99

References 100

TABLES

		Pages
3.1	Definitions of 25 Probation Officer Worker Characteristics	10-12
3.2	37 Task Categories for Probation Officer	14-16
3.3	Average Importance Ratings of 25 Worker Characteristics for 37 Task Categories	17-18
3.4	Reliabilities and Ratings Summary for Probation Officer	19
3.5	Average Usefulness Ratings of 72 Tests as Measures of 17 Before-Hire Worker Characteristics	22-30
3.6	Reliabilities and Ratings Summary of Test to Worker Characteristics Ratings	32
3.7	Linkage Criteria and Summary for Test to Worker Characteristic and Worker Characteristic to Task Category Ratings	33
3.8	Cognitive and Noncognitive Tests for the Experimental Selection Battery	34-35
3.9	Pretest Results and Item Statistics for the Experimental Selection Battery	39
4.1	Probation Officer Performance Dimensions Used During Retranslation	48-50
5.1	Composition of the Final Validation Sample	57
5.2	Gender and Ethnic Status of Final Validation Sample	59
5.3	Education and Tenure of Final Validation Sample	60
5.4	Relationship of Incumbents to Supervisors in Final Validation Sample	61
6.1	Descriptive Statistics for 32 Job Performance Ratings Scales	65
6.2	Correlation of 37 Job Performance Rating Scales with the Overall Criterion	68

TABLES (continued)

		Page
6.3	Factor Descriptions of Five Specific Job Performance Criteria	70
6.4	Descriptive Statistics and Intercorrelations for Five Factor Score Criteria	73
6.5	Descriptive Statistics for the Overall and Specific Job Performance Criteria	74
6.6	Average Overall Criterion Scores by Supervisor and Incumbent Gender	75
6.7	Average Overall Criterion Scores by Supervisor and Incumbent Ethnic Status	78
6.8	Final Item Statistics for the Experimental Selection Battery	79
6.9	Descriptive Statistics on 15 Individual Tests for the Final Validation Sample	80
6.10	Intercorrelations Among Individual Tests	81
6.11	Descriptive Statistics on Individual Tests by Incumbent Gender	82
6.12	Descriptive Statistics on 15 Individual Tests by Incumbent Ethnic Status	83
6.13	Correlations of 15 Individual Tests with Overall Job Performance	86
6.14	Source of Items Selected to Augment the Final Selection Battery	90
6.15	Gender and Ethnic Group Statistics for Probation Officer Examination Scores	91
6.16	Correlations of Individual and Final Predictors with the Five Factor Criterion Measures	92
7.1	Description of the Noncognitive Tests Contributing Items to the Probation Officer Examination	95
7.2	Probation Officer Examination Score Distribution	98

FIGURES

		Page
3.1	Hypothetical Example of the Relationships Among Tests, KSAs, and Job Tasks	8
3.2	Test Development Plan for the Numerical Ability Test	37
4.1	Sample Performance Incident Form	47
4.2	Sample Behavioral Summary Scale	52
6.1	Factor Loadings of 31 Ratings Scales on Five Specific Criteria Factors	71

CHAPTER ONE: INTRODUCTION

The research, analysis, and results described in this report represent the second phase of a three-phase research study sponsored by the Standards and Training for Corrections Program (STC) of the California Board of Corrections. This introduction provides general information about the project and describes the strategy used to conduct the research.

Project Background

The California Board of Corrections had been charged by the State Legislature to undertake several previous studies of the probation (1964) and correctional systems (1971 & 1979). These studies found widely varying hiring standards and training practices from agency to agency--even where printed job descriptions were quite similar. Additionally, many agencies were finding themselves in court faced with the task of defending their hiring procedures while incurring the considerable expense associated with such litigation and facing substantial delays in filling vacancies.

To address these problems, the Board of Corrections made two recommendations to the Legislature. The Board recommended that:

1. A state agency assume responsibility for an analysis and synthesis of job titles and job descriptions and for validation studies of selection standards for corrections personnel.

2. A commission be established, modeled after POST, specifically for standards setting and training for corrections personnel.

In response to the Board's recommendations, the California Probation, Parole, and Correctional Association (CPPCA) proposed Senate Bill 924 which became effective July 1, 1980. SB 924 created the Standards and Training for Corrections Program (STC) to establish minimum standards for the selection and training of local corrections and probation personnel, to certify training courses, and to administer the corrections training fund to help agencies comply with the training standards. Funds for this program are derived from penalty assessments on specified criminal offenses. Participation by individual agencies in the STC program is voluntary. Only those agencies accepting state funds through STC are required to adhere to the selection and training standards.

Subsequently, Assembly Bill 3296 was passed. This bill mandated the Board of Corrections to contract (among other things) for research on validated selection standards for corrections and probation personnel. This mandate provided the impetus for the research efforts described in this report and in related documents.

Participating Departments

County sheriff's or probation departments participating in the STC program were eligible to participate in this project. The 37 eligible sheriff's and 52 eligible probation departments were invited to attend one of three briefings where the goals and procedures of the study were presented. Approximately 20 sheriff's and 21 probation departments attended these briefings.

Departments unable to send representatives to the briefings were mailed copies of the briefing materials and participation agreement forms. Initially, 23 sheriff's departments and 30 probation departments (representing 62% and 58% of eligible departments) elected to participate in the study. The number varied from time to time as departments withdrew or were added to the study.

Research Strategy and Rationale

The research program has consisted of three phases, each corresponding to one of three objectives of the Standards and Training for Corrections Program. These three objectives are:

Phase 1	--	a statewide job analysis of three entry-level positions in county agencies: corrections officer, probation officer, and juvenile counselor.
Phase 2	--	the development and validation of entry-level selection procedures for these positions.
Phase 3	--	the development of training standards and curricula for these positions.

This report focuses primarily upon the second phase of the research project, namely the development and validation of entry-level selection procedures. It is important to note that this report concerns the probation officer position only. Separate reports describe the parallel efforts undertaken for the juvenile counselor and corrections officer positions also included in this project. Of necessity, this report refers frequently to procedures and analyses conducted during the job analysis phase of the project, but is not intended to be a complete description of that research which is described fully elsewhere.[1]

[1]Grabow, K.M., Sevy, B.A., Houston, J.S. Statewide Job Analysis of Three Entry-level Corrections Positions, Personnel Decisions, Inc., 1983. (Available from STC.)

In addition to this report, information about the methods and results of the phase two research for probation officers may be found in the User's Manual as well as in the Candidate Information Booklet and, to a lesser degree, in the test booklet. To maintain test security, access to the test booklets is strictly controlled and closely monitored.

A concurrent, criterion-related strategy was used to guide the development and validation of selection procedures for the entry-level probation officer position. The basic goal of the selections standards phase of the project was to devise a method of accurately predicting the job performance of entry-level probation officers. The criterion-related approach to meeting this goal consisted of developing measures of ability and personal characteristics (predictors) which could be shown to be statistically and meaningfully related to job performance. In addition, the predictors had to be fair and not impose heavy administrative burdens to implement. A concurrent study is one in which predictor and criterion data are collected from job incumbents.

Professional standards (APA, 1980) suggest that a criterion-related approach is preferable whenever it is feasible and provide guidelines for determining the feasibility of such an approach. The first and, for this study, most important guideline concerns the stability of the job. Criterion-related studies are not feasible if the position being studied is in a period of rapid flux. Professional standards are most directly concerned with the stability of a position over time but, due to the variety of departments providing incumbent probations officers to this study, stability of the position across agencies must also be considered. With respect to stability across agencies, job analysis data were analyzed to determine the extent to which the job (as described by probation officer incumbents and supervisors) varied as a function of several parameters (e.g., county size). The results of these analyses clearly demonstrated the equivalence of the job from department to department, thus indicating that the stability of the job across agencies cannot be questioned.

The question of the stability of the probation officer position over time cannot be as directly addressed, but it seems clear that the position is not in a "period of rapid evolution". Nevertheless, the data in this report reflect only a two-and-one-half year "snapshot" of the probation officer position. Should the job undergo significant changes in the future, it will be incumbent upon users of the selection procedures developed in this project to ensure that the procedures remain valid, fair, and useful.

Other guidelines relevant to determining the feasibility of a criterion-related strategy concern the potential for developing a relevant, reliable measure of job performance (criterion); and for collecting predictor and criterion data from a sample of persons that is "reasonably representative of the populations of peoples and jobs to which the results are to be generalized" (APA, 1980).

It is also important that this sample be large enough to allow the reliable detection of meaningful relationships between the predictors and criteria. The research described in this report fully met each of these guidelines. Therefore, a criterion-related approach to the development and validation of entry-level selection standards for probation officers was completely feasible.

Report Overview

The remainder of this report describes the methods used to develop and validate entry-level selection procedures. This description begins in Chapter Two with a summary of the job analysis results that formed the foundation for predictor and criterion development. Development of the predictors and criteria are described separately in Chapters Three and Four. Chapter Five describes the methods used to collect predictor and criterion data from a sample of approximately 300 probation officer incumbents and their supervisors. Chapter Six reports the methods used to analyze the predictor and criterion data and documents the results of those analyses. Finally, Chapter Seven reviews the results with an eye toward their meaning and implications for use of the Probation Officer Examination as part of the pre-employment selection process of participating agencies.

CHAPTER TWO: SUMMARY OF THE JOB ANALYSIS

The job analysis of the probation officer position began with a review of job descriptions, training and orientation materials, and performance evaluation forms supplied by participating probation departments. Document review was followed by site visits to eight counties. During the site visits, facilities were toured and probation officer incumbents and supervisors interviewed individually and in groups to determine the major duties and tasks involved in the probation officer position and the important requirements of the job.

Information gathered from the document review and site visits was used to develop a preliminary version of a job analysis questionnaire (JAQ). The preliminary questionnaire was reviewed by STC staff and by supervisors, training officers, and others from a sample of participating departments. Based on the reviewers' comments, the questionnaire was revised and then pilot tested on a sample of probation officer incumbents and supervisors. During the pilot test, the incumbents and supervisors completed the preliminary questionnaire and provided feedback about its clarity and comprehensiveness.

Based upon the data collected in the pilot test, final revisions of the job analysis questionnaire were made. Two forms of this questionnaire were developed, one for probation officer incumbents and the other for their supervisors. Each questionnaire consisted of four major sections: (1) a general section, collecting information about the person completing the questionnaire; (2) a task rating section where incumbents rated how often they performed each of 245 job tasks and time typically required to complete each task, and supervisors rated the importance of each of those tasks; (3) an equipment rating section in which each of 36 pieces of equipment were rated with respect to their frequency of use on the job; and (4) a worker characteristics section in which each of 98 knowledges, skills, abilities, and personal characteristics were rated in terms of their usefulness for successful job performance.

The final job analysis questionnaires were administered to a random sample of 635 experienced incumbents (those on the job at least six months) and most of the supervisors in the participating counties (274 supervisors). In addition, 84 incumbents and 70 supervisors completed the questionnaire a second time in order to determine its test-retest reliability.

Analysis of the completed questionnaires resulted in a statewide definition of the probation officer "core job", as well as specific definitions of how the job is performed in each participating county

county. The core job was defined, based upon the task ratings from all participating counties, as consisting of all job tasks listed in the job analysis questionnaire which:

1. were performed by half or more of the incumbents across most counties.

OR

2. were rated as having medium or higher importance by the supervisors in most counties.

Both criteria were computed using a procedure that weighted all departments equally. This was done to ensure that the results would be equally applicable to all participating departments and so that the results from large departments did not completely dominate the study. Job tasks not meeting one of the above criteria were excluded from the core job definition -- as were those tasks with medium or higher importance but which were rarely performed in most departments.

Apart from defining the core job for the probation officer position, the job analysis data showed that the job analysis questionnaire was highly reliable over time and that the probation officer job did not vary from county to county. This latter finding was very important as it permitted the development of a uniform set of selection procedures to be applied to agencies throughout the State of California.

For a more detailed description of the research methodology involved in the job analysis, see the Job Analysis Technical Report, December, 1983 (available from the Board of Corrections).

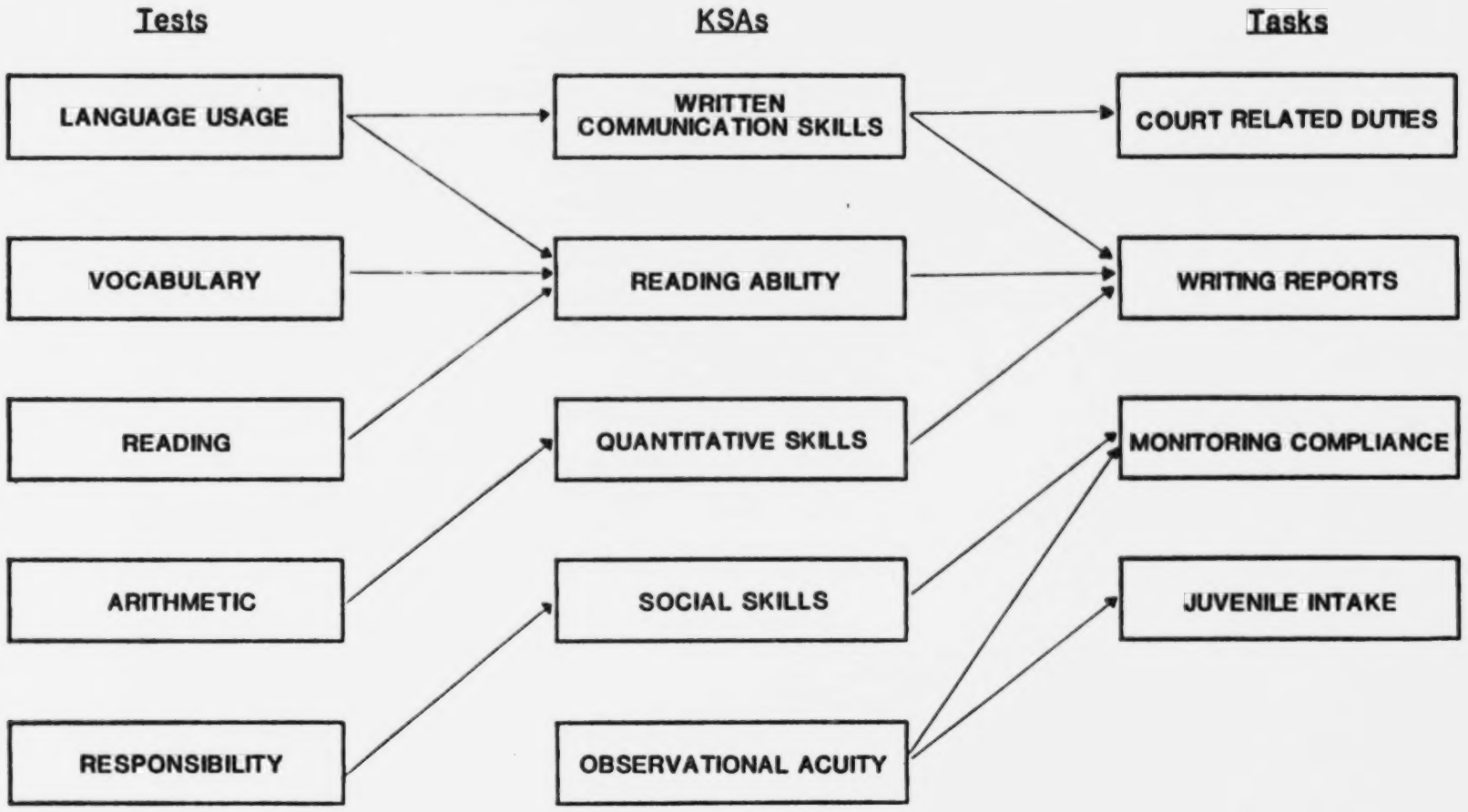
CHAPTER THREE: PREDICTOR DEVELOPMENT

Overview and Rationale

The end goal of predictor development was to arrive at a set of tests (experimental selection battery) to be administered to a sample of probation officer incumbents (validation sample). These tests would comprise the building blocks from which the Probation Officer Examination would be constructed. Because the tests included in the experimental selection examination would form the basis of the final, implemented examination, it was very important that they be highly indicative of candidates' ability to perform all important job tasks. It was necessary that the tests predict performance on the tasks that make up the job. However, tests do not measure task performance directly. What tests do measure are knowledges (e.g., knowledge of departmental rules and regulations), skills (e.g., skill in listening), abilities (e.g., ability to recognize small details), or personal characteristics (e.g., integrity). In order to link the tests to the job tasks, it was necessary to use the knowledges, skills, and abilities (KSAs) as intermediaries. Tests measure KSAs, not job tasks. Once it could be shown which KSAs are essential to the performance of which tasks, tests could be selected that measure the KSAs and, therefore, that predict task performance. Figure 3.1 shows the hypothetical relationships among a set of tests, KSAs, and tasks. The arrows show which tests are useful measures of which KSAs and which KSAs are required for the successful performance of specific tasks. Successful report writing, for example, requires three KSAs (Written Communications Skills, Reading Ability, and Quantitative Skills). Four tests are useful measures of one or more of these three KSAs. Because those tests are useful measures of the set of KSAs necessary for successful report writing, they can be used to predict report writing performance. In this project, the determination of which KSAs are important for specific tasks was made by probation officer supervisors who rated the importance of every KSA for the successful performance of every job task. The usefulness of each of 72 tests as a measure of every KSA was rated by PDI staff psychologists experienced in the use of tests in making selection decisions.

Chapter 3 describes the methods used to link tasks and tests to KSAs. Also in Chapter 3 is a description of how several new tests were drafted, pretested, and refined for inclusion in the experimental selection battery.

Figure 3.1 Hypothetical Example of the Relationships Among Tests, KSAs, and Job Tasks



Linking KSAs to Job Tasks

As was noted, one aspect of predictor development involved the linking of KSAs (directly measurable) to job tasks (ultimate criterion). Linking was accomplished by having job experts (probation officer supervisors) rate the importance of each knowledge, skill, or ability for successful performance of the job tasks. Because there were 245 job tasks and 98 KSA items in the job analysis questionnaire, a reduction in the number of elements to be rated was necessary if the task was to be administratively feasible.

Reducing the number of KSAs to be rated was accomplished by combining them into broader groupings. PDI staff members rationally clustered the 98 individual KSAs into groups representing common characteristics (see Table 3.1). For example, several KSA items concerning Department-specific knowledge were grouped together. In this way, the 98 KSAs were combined into 25 worker characteristic clusters. This achieved a four-fold reduction in the number of KSAs to be rated, and at the same time, preserved the variety of skills and abilities important for successful job performance. The resulting worker characteristic clusters were reviewed by other PDI staff members and then "confirmed" via corroborative factor analysis.

Confirmation consisted of two separate oblique factorings of the worker characteristic data: one using incumbents' ratings of the usefulness of each KSA for successful job performance, and one using similar ratings from supervisors. The corroborative factor analysis began with the assignment of every KSA to a target factor (i.e., one of the 25 worker characteristic clusters). Then, based upon the intercorrelations of the KSAs, the appropriate matrix algebra equations were solved yielding a new matrix that showed the relationships among the KSAs and factors (Overall & Klett, 1971). This matrix was examined to corroborate the rational clustering of the KSAs. The clustering was corroborated when the KSAs correlated more highly with their target factor than they did with any of the other factors. Generally, the worker characteristics were corroborated, but the results suggested that a few KSAs were more strongly related to other clusters. So long as the relocations made sense, KSAs were relocated to different clusters and the correlations recomputed to corroborate the new structure.

As noted, the corroboration was done separately using the JAQ data of incumbents and supervisors. It was found that the KSA items migrating, for example, from worker characteristic "A" to worker characteristic "F" when incumbents' data were analyzed, had a high probability of making the same migration when the supervisors' data were analyzed. This indicated that the rational clustering of KSAs into worker characteristics and subsequent minor revisions provided reliable results.

Table 3.1

Definitions of 25 Probation Officer Worker Characteristics

A. Department-Specific Knowledge Knowledge of departmental rules, regulations, procedures, and available community resources.

B. Knowledge of Specific Laws Knowledge of codes (e.g., Penal, Welfare and Institutions, Government, Civil, Probate, Health and Safety, Business and Professions, and Vehicle Codes), of adult and juvenile laws and procedures, and of laws of search and seizure.

C. Knowledge of the Legal System Terminology Knowledge of legal terminology and documents, civil and court procedures, the criminal justice system, the confidentiality of records (C.O.R.I.), the rights of probationers, and the potential legal liabilities of probation officers.

D. Knowledge of Psychology Knowledge of: symptoms of emotional disorders and mental illness; human development; counseling, crisis intervention, and problem solving techniques; and value systems of different groups or cultures.

E. Knowledge of Chemical Use and Abuse Knowledge of the symptoms and effects of drug and alcohol use and abuse, and of drug and alcohol testing procedures.

F. Interviewing and Counseling Skill Skill in gathering information through interviews, in listening, and in counseling others (cooperative and uncooperative individuals, adults and children, including abused or neglected children) to provide support and facilitate behavior change.

G. Safety and First Aid Knowledge of safety and emergency procedures, and skill in administering first aid and CPR.

H. Oral Communication Skill in communicating verbally.

I. Facility with Another Language Skill in speaking and understanding Spanish or Asian languages.

J. Written Communication Knowledge of grammar, spelling, punctuation, and sentence structure. Skill in writing legibly, and in preparing clear and understandable reports, case notes, and other correspondence.

Table 3.1 (continued)

K. Quantitative Skill Knowledge of basic mathematics (add, subtract, multiply, and divide).

L. Reading Skill in reading sufficiently well to understand laws, statutory codes, and court and other legal documents.

M. Skill in Observing Behavior Skill in accurately observing and documenting behavior, and in remaining alert to signs of potential problems on the job.

N. Searching Skill in searching persons, vehicles, and residences or institutional facilities for contraband or weapons; ability to notice small details when searching or doing surveillance.

O. Self-Defense/Physical Restrain of Others Skill in using self-defense, physical restraint techniques, and restraining gear (such as handcuffs).

P. Social Skills Skill in public relations; effectiveness in working with others and interacting with diverse groups; consistency and fairness in dealing with others; insight, sensitivity, and concern for others' feelings; tolerance for differences in values and beliefs; sense of humor.

Q. Motivation/Dependability Dependability and trustworthiness (e.g., coming to work on time, not abusing sick leave, assuming a fair share of the workload); willingness to put in extra time and effort; persistence in seeking information; integrity in adhering to professional ethics; commitment to the facility's goals and objectives.

R. Stress Tolerance Tolerance for the stress of working in a potentially dangerous environment; ability to maintain self-control when confronted with verbal harassment or threats; ability to stay calm and respond with composure in emergencies.

S. Dominance/Self-Esteem Assertiveness, self-assurance, and decisiveness; ability to influence and confront others, to enforce rules, to interact confidently with high-status individuals (e.g., judges).

T. Intelligence and Judgement Ability to effectively recall, analyze, and synthesize large amounts of information, and to make sound independent decisions where there is no standard procedure.

Table 3.1 (continued)

U. Personal Organization Skill in organizing one's work, files, and other materials, and efficiency in use of time, meeting deadlines, and working effectively under time pressure.

V. Tolerance of Negative Aspects of the Job Tolerance of physically unpleasant work environments; ability to perform repetitive tasks effectively or tolerate long periods of inactivity.

W. Flexibility Resourcefulness in thinking of alternative approaches (e.g., alternative treatment techniques, referral sources, etc., ability to adjust to new and unfamiliar situations/duties, resilience in coping with setbacks.)

X. Visual Acuity Daytime and nighttime visual acuity (ability to see well under a variety of light conditions) and depth perception.

Y. Physical Abilities Upper body and leg strength; whole body coordination; manual dexterity and eye/hand coordination; endurance/stamina; and agility/flexibility.

Grouping Task Items into Task Categories

In the job analysis questionnaire, the 245 individual task items had been grouped into 19 broad categories of job responsibility (e.g. Preparing Court Reports). For the same reason that the KSAs were grouped into a fewer number of worker characteristics, it was necessary to group the task items into a fewer number of task categories. This reduction was accomplished by rewriting the individual task items in each general JAQ category into one to six intermediate categories, each consisting of task items related to similar behaviors or responsibilities (e.g., Reviewing charges, previous record, see Table 3.2). The resulting 37 task categories (a six-fold reduction in number) were reviewed for completeness and were then used to link job tasks to the worker characteristics.

Task Category to Worker Characteristic Ratings

Fifteen probation officer supervisors from four agencies rated how important the 25 worker characteristics were for the successful performance of each of the 37 task categories (approximately 900 ratings per supervisor). On average, the raters were responsible for supervising 7 probation officers, and had been supervisors for an average of 8 years. They had themselves served as probation officers for an average of 7 years. Ratings were made using a 5-point scale, indicating the importance of each worker characteristic for every task category (1 = not important, 2 = low, 3 = medium, 4 = high, and 5 = very high importance). Table 3.3 contains the average importance rating of each worker characteristic for every task category.

The reliability of these ratings was estimated using the intraclass correlation approach described by Shrout and Fleiss (1979). Twenty-five reliability estimates were computed, one for each worker characteristic rated. The ratings were reliable to the extent that the within-task category ratings variance was significantly less than the between-task categories variance. The specific model used to estimate the reliability of the ratings treats raters as a fixed effect (i.e., did not seek to generalize the results beyond those particular raters) and used the mean rating as the unit of analysis. The analysis sought only to estimate the reliability of the ratings as they actually were used in the study. Because the average rating was used to link worker characteristics to task categories, the reliability of the average rating, rather than the raters per se, was estimated.

The reliability of the ratings for each worker characteristic are shown in Table 3.4. Reliabilities range from .94 (Self-Defense) to .56 (Motivation) with an average of .83, indicating that the ratings were made very reliably and therefore it is unlikely that the average rating would have differed appreciably should the supervisors have made the ratings a second time.

Table 3.2

37 Task Categories for Probation Officers

Juvenile Intake and Processing.

1. Reviewing charges, previous record.
2. Notifying parents; interviewing, counseling, and advising juvenile and parents of rights, informing of hearing date.
3. Making recommendations and decisions concerning wardship, detention, release.

Adult or Juvenile Investigation (including Violation of Probation).

4. Interviewing juvenile/adult, co-defendants, witnesses, victims, and officer about offense.
5. Reviewing information about current report.
6. Reviewing past records, obtaining medical, psychiatric, employment, education, and past treatment information from records and other people.
7. Contacting victim about court appearance rights and deadlines.

Making Recommendations for Disposition of New Cases or Violations of Probation.

8. Evaluating seriousness of offense, aggravating or mitigating circumstances, (based on investigation), and recommending sentence and/or conditions of probation.
9. Recommending placement and making arrangements with facility.

Establishing Supervising Relationship with New Probationer. (Juvenile or Adult).

10. Arranging and conducting initial interview, to assess probationer's needs, and the risks posed to the community.
11. Reviewing probationer's file, consulting with outside agencies to gain familiarity with case.
12. Establishing treatment plan, and reviewing conditions and terms with probationer and family.

Table 3.2 (continued)

Setting Up Collection of Restitution, Fines and Other Payments.

13. Obtaining and evaluating appropriateness of victims' restitution claims.
14. Recommending amount due and repayment schedule based on probationer's ability to pay.

Monitoring Compliance of Adult/Juvenile with Terms of Probation.

15. Interviewing probationer and others (parents, employers) to check on compliance with terms and conditions.
16. Visiting residence, work place.
17. Corresponding with referral sources, monitoring restitutions.
18. Taking action due to lack of progress (arresting, requesting warrant, changing frequency of contact, initiating court process).
19. Writing reports for violation of probation, request for supplemental action.
20. Reviewing request to leave state or county.

Making Referrals to Providers of Medical, Mental Health, Social Service, Education, Employment, and Other Service.

21. Determining needs (for treatment, evaluation, etc.)
22. Referring probationer or family members to the appropriate program.
23. Following-up to verify and evaluate success.

Informing, Assisting, Advising, and Counseling.

24. Conveying relevant information (law violations, change of placement) about probationer to police department, agency (e.g., welfare, loans).
25. Helping probationer with financial, vocational, or personal counseling, answering questions.
26. Intervening in verbally explosive and physically violent situations.

Preparing Court Reports.

27. Preparing pre-sentence, disposition or supplemental reports for (Municipal, Superior, or Juvenile) Court.

Table 3.2 (continued)

Other Paperwork.

28. Writing or dictating correspondence, reports, taking notes, keeping records.
29. Gathering and organizing information from files, proofreading, filing.

Other Court-related Duties.

30. Reviewing case before court appearance.
31. Testifying and answering questions in court, making recommendations and providing oral updates.
32. Conducting investigations and studies ordered by the court.

Providing Service to the Community.

33. Giving presentations, responding to questions from public, speaking with juveniles at parents request.

Establishing and Maintaining Relationships with Referral and Placement Resources.

34. Developing and maintaining contacts with foster parents, potential employers, placement facilities, and social service agencies.

Miscellaneous Duties.

35. Keeping informed about department via meetings, training courses, consultation with co-workers and supervisor.
36. Transporting and arranging transportation of individuals.
37. Giving assignments to, training and assisting other personnel.

BEST COPY AVAILABLE

Table 3.3 Average Importance Ratings of 25 Worker Characteristics for 37 Task Categories

Rating Scale:
1 = NOT IMPORTANT
2 = LOW IMPORTANCE
3 = MEDIUM IMPORTANCE
4 = HIGH IMPORTANCE
5 = VERY HIGH IMPORTANCE

JOB TASK CATEGORIES	Task No.	A Department-Specific Knowledge	B Knowledge of Specific Laws	C Knowledge of the Legal System & Terminology	D Knowledge of Psychology	E Knowledge of Chemical Use & Abuse	F Interviewing & Counseling Skill	G Safety & First Aid	H Oral Communication	I Facility with Another Language	J Written Communication	K Quantitative Skill	L Reading	M Skill in Observing Behavior	N Searching	O Self-Defense/Physical Restraint of Others	P Social Skills	Q Motivation/Dependability	R Stress Tolerance	S Dominance/Self-Esteem	T Intelligence & Judgement	U Personal Organization	V Tolerance of Negative Aspects of the Job	W Flexibility	X Visual Acuity	Y Physical Abilities
I. Juvenile Intake and Processing	1	3.1	4.3	4.1	3.1	2.1	2.1	1.2	2.6	1.3	2.4	2.1	4.4	1.8	1.4	1.3	1.9	3.1	2.2	1.9	3.8	3.1	2.1	2.4	2.1	1.4
	2	3.9	4.3	4.0	3.1	2.5	4.6	1.2	4.4	2.5	2.7	1.8	3.2	3.5	1.6	1.6	3.9	3.3	3.0	3.1	3.7	3.2	2.6	3.2	1.9	1.7
	3	4.3	4.3	4.3	3.1	2.5	2.8	1.1	3.5	1.8	4.4	2.0	3.5	2.9	1.3	1.3	2.9	3.1	2.8	2.9	4.5	3.3	2.4	3.4	1.8	1.4
II. Adult or Juvenile Investigation (including Violation of Probation)	4	3.8	4.3	4.3	3.4	2.8	4.6	1.3	4.6	2.6	2.9	1.8	3.1	4.3	1.4	1.5	3.8	3.3	3.5	3.4	4.3	3.2	2.8	3.1	2.1	1.8
	5	3.8	3.7	4.0	2.3	2.4	1.8	1.1	1.9	1.4	2.3	2.0	4.6	1.8	1.3	1.3	1.9	2.9	2.0	1.9	4.3	3.4	2.1	2.3	2.4	1.3
	6	3.6	3.7	3.8	3.3	2.7	2.9	1.1	3.7	2.2	3.3	1.9	4.6	2.1	1.3	1.3	3.2	3.4	2.1	2.3	4.0	3.3	2.1	2.8	2.4	1.5
	7	3.6	3.9	3.9	2.8	1.6	3.3	1.1	4.4	2.8	2.8	1.6	2.6	2.4	1.2	1.3	4.0	3.1	2.6	2.9	3.2	3.1	2.3	2.6	1.4	1.4
III. Making Recommendations of Disposition of New Cases or Violations of Probation	8	4.1	4.6	4.4	3.3	2.8	2.9	1.1	3.2	1.8	3.8	2.4	4.3	2.2	1.3	1.2	2.8	3.2	2.4	2.6	4.4	3.8	2.7	3.5	1.8	1.4
	9	4.3	4.2	4.0	3.5	2.8	3.1	1.2	4.0	2.1	3.4	1.9	2.7	3.2	1.3	1.3	3.4	3.4	2.3	3.1	4.2	3.4	2.3	3.5	1.6	1.5
IV. Establishing Supervising Relationship with New Probationer (Juvenile or Adult)	10	3.8	3.9	3.9	4.1	3.5	4.8	1.3	4.6	2.6	2.4	1.8	2.5	4.1	1.4	1.5	3.9	3.1	3.1	3.4	4.4	3.3	2.6	3.5	1.8	1.6
	11	3.4	3.5	3.7	3.2	2.5	3.3	1.1	3.9	2.2	3.1	1.7	3.9	2.3	1.3	1.2	3.7	3.3	2.4	2.7	3.8	3.5	2.3	3.0	1.6	1.5
	12	3.8	3.8	3.7	3.9	2.8	4.4	1.2	4.5	2.5	2.8	1.6	2.8	3.6	1.7	1.6	4.0	3.3	3.1	3.3	4.1	3.4	2.5	3.3	1.7	1.6
V. Getting Up Collection of Restitution Fines, and Other Payments	13	3.9	3.4	3.4	2.4	1.3	2.9	1.1	3.3	1.9	3.1	3.4	3.5	2.1	1.3	1.1	3.1	2.9	2.4	2.6	3.9	3.3	2.6	2.7	1.6	1.4
	14	3.8	3.4	3.5	2.0	1.4	2.8	1.2	3.6	1.9	3.3	3.8	3.0	1.8	1.4	1.1	2.9	2.9	2.5	2.9	3.7	3.2	2.6	2.7	1.6	1.3
VI. Monitoring Compliance of Adult/Juvenile with Terms of Probation (or Workshop/Dependency	15	4.1	4.0	3.8	3.4	2.8	4.3	1.2	4.6	2.5	2.5	1.8	2.4	4.1	2.2	1.8	3.9	3.5	3.3	3.5	4.0	3.4	2.8	3.5	1.8	1.5
	16	3.4	2.8	2.5	3.0	2.8	3.6	1.7	4.0	2.5	1.8	1.6	1.9	4.0	2.6	2.3	3.9	3.4	3.1	3.4	4.1	3.1	3.0	3.3	2.1	1.8
	17	3.6	3.2	3.3	1.8	1.8	2.3	1.1	2.9	1.9	4.4	3.1	3.5	1.8	1.3	1.2	2.6	3.1	1.9	2.8	3.5	3.4	2.6	2.8	1.6	1.3
	18	4.5	4.4	4.4	3.3	3.1	3.9	1.9	4.0	2.3	3.9	2.0	2.9	3.5	2.8	2.9	3.7	3.4	3.6	3.6	4.4	3.4	2.8	3.4	2.1	2.7
	19	4.4	4.3	4.4	2.6	2.3	2.6	1.1	2.7	1.8	4.9	2.2	3.4	2.9	1.4	1.2	2.9	3.3	2.6	2.6	4.1	3.6	2.4	3.4	1.7	1.4
	20	4.5	4.0	4.1	2.1	1.6	2.3	1.1	2.8	1.7	2.8	1.8	3.3	2.5	1.3	1.1	2.8	2.9	2.1	2.3	3.5	2.9	1.8	2.8	1.6	1.4
VII. Making Referrals to Providers of Medical, Mental, Health, Social Service, Education, Employment, Other Service	21	4.1	3.8	3.8	3.9	3.3	4.2	1.1	3.4	2.1	2.6	1.8	3.4	3.9	1.3	1.1	3.2	3.3	2.4	2.5	4.3	3.1	1.9	3.4	1.6	1.3
	22	3.9	3.3	3.3	3.9	3.1	3.8	1.1	3.9	2.3	3.3	1.9	2.6	3.3	1.3	1.1	3.6	3.3	2.7	3.1	4.1	3.3	2.4	3.4	1.7	1.4
	23	3.2	3.4	2.6	2.8	2.8	3.1	1.2	3.6	2.1	3.5	1.7	2.9	3.0	1.4	1.3	3.2	3.1	2.4	2.9	4.1	3.3	2.3	3.0	1.7	1.4
VIII. Informing Assisting, Advising, and Counseling	24	4.3	4.0	4.1	2.5	2.3	2.7	1.4	4.3	1.6	3.7	1.8	2.6	2.3	1.4	1.3	3.5	3.1	2.3	3.1	3.9	3.1	2.3	2.8	1.4	1.3
	25	3.8	3.3	3.3	3.1	2.6	4.4	1.4	4.4	2.4	2.8	2.0	2.9	3.4	1.4	1.6	4.0	3.1	2.8	3.2	4.1	3.1	2.5	3.6	1.8	1.4
	26	3.9	3.6	3.1	4.5	2.9	4.3	2.8	4.3	2.7	1.6	1.6	1.5	4.2	2.5	3.9	4.0	3.2	4.1	3.8	4.4	2.3	3.3	3.9	2.5	3.6
IX. Preparing Court Reports	27	4.5	4.4	4.6	3.3	3.1	3.4	1.4	2.8	1.9	4.9	2.4	3.7	2.6	1.4	1.3	3.1	3.6	2.5	3.2	4.3	4.0	2.3	3.6	1.8	1.4
X. Other Paperwork	28	3.6	3.0	3.8	2.7	2.5	2.2	1.3	3.6	1.6	4.3	2.3	3.1	1.9	1.3	1.3	2.6	3.4	2.2	2.4	3.9	4.1	2.5	3.2	1.8	1.4
	29	3.1	2.6	2.9	2.0	1.9	1.6	1.3	1.8	1.4	2.7	2.1	4.1	1.6	1.3	1.1	1.9	3.1	2.1	2.1	3.4	4.4	2.9	2.9	1.8	1.4
XI. Other Court Related Duties	30	3.4	4.0	4.1	2.4	2.4	2.0	1.2	2.1	1.5	2.3	1.8	4.5	1.9	1.3	1.1	1.9	2.6	2.2	2.3	3.8	3.4	2.0	2.4	2.0	1.5
	31	4.3	4.3	4.7	3.0	2.9	2.5	1.1	4.9	1.4	2.2	1.9	2.6	2.6	1.3	1.1	3.6	3.3	3.6	3.9	4.2	3.4	2.6	3.0	1.6	1.5
	32	4.3	4.3	4.4	3.3	3.0	4.3	1.3	3.8	2.0	4.1	2.1	3.4	3.7	1.7	1.4	3.8	3.6	3.0	3.6	4.1	3.8	2.4	3.4	1.9	1.6
XII. Providing Service to the Community	33	4.3	4.3	4.1	3.4	3.1	3.4	1.2	4.6	2.1	2.4	1.7	2.6	2.9	1.3	1.3	4.3	3.6	3.2	3.8	4.2	3.6	2.4	3.1	1.7	1.4

Table 3.3 (continued)

Rating Scale:

1 = NOT IMPORTANT
2 = LOW IMPORTANCE
3 = MEDIUM IMPORTANCE
4 = HIGH IMPORTANCE
5 = VERY HIGH IMPORTANCE

JOB TASK CATEGORIES	Task No.	A Department-Specific Knowledge	B Knowledge of Specific Laws	C Knowledge of the Legal System & Terminology	D Knowledge of Psychology	E Knowledge of Chemical Use & Abuse	F Interviewing & Counseling Skill	G Safety & First Aid	H Oral Communication	I Facility with Another Language	J Written Communication	K Quantitative Skill	L Reading	M Skill in Observing Behavior	N Searching	O Self-Defense/Physical Restraint of Others	P Social Skills	Q Motivation/Dependability	R Stress Tolerance	S Dominance/Self-Esteem	T Intelligence & Judgment	U Personal Organization	V Tolerance of Negative Aspects of the Job	W Flexibility	X Visual Acuity	Y Physical Abilities
XIII. Establishing and Maintaining Relationships with Referral and Placement Resources	34	3.8	3.4	3.2	3.4	2.6	3.4	1.3	4.4	1.9	3.2	1.9	2.9	3.0	1.4	1.4	3.9	3.4	2.8	3.4	3.9	3.6	2.4	3.4	1.7	1.5
XIV. Miscellaneous Duties	35	4.2	3.3	3.4	2.7	1.8	2.3	1.4	4.0	1.3	2.8	1.8	3.8	2.3	1.2	1.2	3.3	3.5	2.3	2.6	3.6	3.4	2.1	2.6	1.6	1.4
	36	4.1	3.3	2.4	3.1	2.4	2.6	2.9	3.3	2.1	2.1	1.7	2.1	3.3	2.9	3.3	3.3	3.1	2.9	3.3	3.4	2.8	2.4	2.6	2.1	2.7
	37	4.4	4.2	4.3	3.5	2.4	3.5	1.6	4.6	1.6	3.8	2.2	3.6	3.3	1.5	1.6	4.1	4.1	2.8	3.8	4.4	4.0	2.9	3.9	1.8	1.6

Table 3.4

Reliabilities and Ratings Summary
for Probation Officers

Worker Characteristic	Rating rel.	Strength of[b] Task Relationships Low	Mod.	High
A. Department specific knowledge	.75	0	7	30
B. Knowledge of specific laws	.83	0	12	25
C. Knowledge of the legal system and terminology	.87	0	11	26
D. Knowledge of psychology[a]	.89	0	30	7
E. Knowledge of chemical use/abuse	.84	2	34	1
F. Interviewing/counseling skill[a]	.92	0	24	13
G. Safety/first aid	.90	32	5	0
H. Oral Communicationa	.91	0	12	25
I. Facility w/other language[a]	.85	5	32	0
J. Written Communication[a]	.92	0	26	11
K. Quantitative Skill[a]	.85	0	36	1
L. Reading[a]	.91	0	24	13
M. Skill in observing behavior[a]	.91	0	27	10
N. Searching	.86	28	9	0
O. Self-defense/restraint of others	.94	27	9	1
P. Social Skills[a]	.90	0	19	18
Q. Motivation/dependability[a]	.56	0	31	6
R. Stress Tolerance[a]	.82	0	33	4
S. Dominance/self-esteem[a]	.86	0	29	8
T. Intelligence/judgment[a]	.68	0	3	34
U. Personal Organization[a]	.78	0	27	10
V. Tolerance of negative job aspects[a]	.58	0	37	0
W. Flesxibility[a]	.78	0	29	8
X. Visual Acuity	.65	2	35	0
Y. Physical Ability[a]	.92	20	16	1

[a]Before-hire worker characteristic

[b]See text (p. 20) for details.

Setting Linkage Criteria

Once the task category to worker characteristic ratings were collected and determined to have been reliable, it was necessary to document the results and to define the level of importance (i.e., average rating) at which a worker characteristic would be considered linked to a task category. For purposes of summarizing the strength of the task category to worker characteristic linkages, the 5-point rating scale was trichotomized into low (average rating of 1.0 - 1.5). moderate (1.5-3.5). and high (3.5 - 5.0). Table 3.4 summarizes the number of task categories linked to each worker characteristic at each of the three linkage strengths. As can be seen, for approximately 90% of the worker characteristics, the task category linkages were primarily of a single strength. For the remaining 10%, the linkages were approximately evenly divided between two strengths.

It was desired to set the criterion for linking a worker characteristic to a task category as high as possible in order that the tests developed to measure each linked worker characteristic would provide an accurate prediction of task performance. Developing a test to measure a worker characteristic that is only weakly linked to the task category would not result in a selection battery likely to provide accurate predictions of job performance. Initially, the criterion was set at 4.0; that is, only those worker characteristics with average importance ratings of 4.0 or higher would be considered linked to a task category. As will be discussed however, it was necessary to relate this criterion to ensure that all of the tasks were linked to at least one predictor. At the time these data were being collected and analyzed, the core job had not been defined (see Job Analysis Report for details). As a result, the relative importance of the 245 job tasks was not known. Rather than risk overlooking a task category that may later have been found to be an important, or even a critical aspect of the probation officer job, it was decided to set the linking criterion so that every task category, therefore every task, would be linked to at least one worker characteristic for which a test could be developed.

Linkage Tests to Worker Characteristics

In much the same way that worker characteristics were linked to task categories, it was necessary to link the tests being considered for inclusion in the experimental selection battery to the same set of worker characteristics (recall that it is through the worker characteristics that tests were linked to the tasks they were to predict). The first step in this process was to compile a list of tests having potential to predict the job performance of probation officers. Based upon research reports located during the literature review, the job analysis materials and results, and the experience of PDI staff members, a list of 72 tests was compiled. For each test a one-page description was constructed describing the characteristic it measured, its reliability and, if available, a summary of its validity for similar jobs.

Eleven psychologists from PDI and its sister organization Personnel Decisions Research Institute then rated the degree to which each of the 72 tests would be a useful measure of each of the probation officer worker characteristics. The full set of 25 worker characteristics was narrowed to 17 by deleting those characteristics which, on the basis of the job analysis data, were found as not being necessary before hire. This was done because any worker characteristic not necessary at the time of hire (i.e., that could be trained on the job) would not be an appropriate selection standard and, therefore, linking tests to those worker characteristics would have served no useful purpose. In order that no worker characteristic be prematurely rejected from consideration, this narrowing of the list of worker characteristics was done as conservatively as possible. Before-hire worker characteristics were defined as those having at least one constituent KSA which, according to at least 40% of probation officer supervisors, should be developed before hire (i.e..at least 40% of supervisors rated the KSA 0 or 1 on the 0 to 4 "when developed" JAQ scale).

Ratings of the usefulness of each test as a measure of each worker characteristic were made on an 8-point scale ranging from "not at all useful" (rating of 0), through "almost useless" (ratings of 1 or 2), "of some use" (ratings of 3, 4, or 5), "highly useful (ratings of 6 or 7), to the worker characteristic being "entirely measured" by the test (rating of 8). Table 3.5 shows the average usefulness of each test as a measure of each of the 17 before-hire worker characteristics.

The reliability of the test to worker characteristic ratings was determined using the same methods described for determining the reliability of the task category to worker characteristic ratings. The reliability analyses indicated that, with the exception of the Physical Ability worker characteristic, the ratings were all quite reliable (range = .72 to .97, average = .86). The reliability of the physical ability ratings (intraclass r = .08) at first glance appears quite dismal, but this apparent unreliability is a statistical artifact rather than an indication of unreliable ratings. The intraclass correlation estimate of reliability is an index of the relative amount of variation in the ratings data. As the data in Table 3.5 reveal, there was almost no variation in the psychologists' ratings of the usefulness of the 72 tests as measures of the Physical Ability worker characteristic (i.e., no tests were rated as useful measures of physical abilities). Because there was essentially no variation in the ratings, the intraclass correlation must be near zero: Nevertheless the ratings were made reliably in the sense that all raters agreed as to the usefulness (lack of usefulness, actually) of the 72 tests as measures of probation officer physical abilities.

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 17 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Interviewing & Counseling Skill	Oral Communication	Facility with Another Language	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Personal Organization	Tolerance of the Job of Negative Aspects	Intelligence & Judgment	Flexibility	Physical Abilities
1. EAS Ver. Comp.	3.3	2.7	5.0	2.7	5.0	5.4	1.4	1.3	1.4	.4	.4	1.2	1.0	.4	5.3	1.4	0
2. EAS Num. Ab.	1.5	.3	.4	.8	1.3	1.2	7.5	.3	.1	.2	.1	.4	.4	.2	3.0	.5	0
3. EAS Visual S & A	.3	.2	.1	.9	1.4	1.0	1.2	2.3	.1	.2	.1	.2	1.4	.9	1.3	.1	0
4. EAS Space Vis.	.1	.1	.2	.2	.4	.3	1.1	1.0	.1	.1	.1	.1	.6	.2	1.1	.1	.2
5. EAS Ver. Reas.	3.4	2.7	3.3	1.4	3.8	3.6	.9	1.1	1.1	.1	.2	.6	.6	.1	5.5	1.0	.1
6. EAS Symb. Reas.	2.1	1.2	1.1	1.4	1.5	1.8	2.0	.6	.3	.1	.1	.4	.4	.1	4.5	1.1	.1
7. DAT Abst. Reas.	2.7	1.6	1.4	1.4	2.0	1.6	2.3	1.2	.4	.1	.1	.1	.4	.1	4.9	1.1	.1
8. DAT Lang. Usage	2.1	1.5	4.2	1.9	4.1	6.3	1.0	.3	.7	.1	0	.4	.4	.1	3.1	.3	.1
9. DAT Ver. Reas.	3.8	3.0	3.6	1.4	3.6	3.8	.9	1.1	.7	.1	.1	.4	.4	.1	5.6	.9	.1
10. EEI Follow D.	.8	.6	1.8	.8	1.7	.8	1.0	.7	.2	.4	.3	.2	1.1	.9	1.4	.3	.1
11. EEI Tables & Gra.	.8	.3	.5	.3	1.4	.6	1.3	.9	.1	.1	.1	.2	.6	.4	1.2	.2	.1
12. ETS Hid. Fig.	.2	.1	.1	.1	.4	.4	.3	2.3	.1	.1	.1	.2	.3	.2	1.2	1.6	.1
13. ETS Card Rotat.	.2	.1	.1	.4	.5	.3	.6	1.9	.1	.1	.1	.1	.3	.1	1.3	1.1	.1

[a]Full Test Names Follow This Table

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 17 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Interviewing & Counseling Skill	Oral Communication	Facility with Another Language	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Personal Organization	Tolerance of the Job of Negative Aspects	Intelligence & Judgment	Flexibility	Physical Abilities
14. ETS Map Mem.	.4	.4	.2	.6	.6	.3	.5	3.1	.1	.1	.1	.1	.4	.1	1.2	.2	.1
15. ETS Bldg. Mem.	.3	.4	.1	.4	.4	.2	.4	3.0	.1	.1	.1	.1	.4	.1	1.0	.2	.1
16. ETS Follow Dir.	.8	.6	2.0	.5	2.2	.9	.7	.7	.1	.1	.1	.1	.6	.8	2.1	.6	.1
17. Watson-Glaser	3.4	2.9	2.9	1.1	4.1	3.2	1.4	1.5	1.2	.3	.1	.6	1.3	.1	6.2	.9	.1
18. CO Reading	2.0	1.3	2.3	1.1	7.0	2.8	1.0	.6	.3	.1	.1	.1	.5	.1	2.7	.3	.1
19. Otis-Lennon	3.2	2.4	2.8	1.7	3.4	3.0	3.4	1.5	1.0	.3	.1	.4	1.1	.1	6.2	1.0	.1
20. Wesman Verb.	3.2	2.4	3.0	2.0	3.9	3.4	1.8	1.4	1.1	.3	.1	.2	1.2	.1	5.8	.9	.1
21. CAB Ma	1.3	.9	.9	1.5	1.4	1.3	.8	2.0	.3	.1	.1	.1	.2	.4	2.0	.2	.1
22. CPI Do	.8	2.4	2.3	.3	.3	.9	.4	1.1	3.2	1.2	1.6	6.6	1.4	1.0	1.5	1.4	.1
23. CPI Sy	1.3	3.2	2.4	.4	.5	.3	.2	1.4	5.3	.8	.8	2.6	.4	.5	.5	1.4	.1
24. CPI Wb	1.2	2.9	1.0	0	.3	.1	.1	.7	3.4	2.8	2.9	3.5	.8	3.0	.3	1.4	.1
25. CPI Re	.9	1.4	.4	0	.3	.4	.1	.7	1.7	5.9	1.9	1.4	2.0	2.7	.8	1.0	.1
26. CPI So	.5	2.0	.4	0	.3	.2	.1	1.4	2.4	5.3	2.0	1.4	1.2	3.2	.8	1.5	.1

[a]Full Test Names Follow This Table

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 17 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Interviewing & Counseling Skill	Oral Communication	Facility with Another Language	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Personal Organization	Tolerance of the Job	Intelligence & Judgment	Flexibility	Physical Abilities
27. CPI Sc	1.3	1.9	1.0	0	.3	.2	.1	.5	2.4	3.2	4.3	1.6	1.8	2.3	.9	1.6	.1
28. CPI To	1.7	3.6	1.5	.3	.4	.3	.3	1.4	4.4	1.7	2.2	1.7	.6	2.2	1.4	2.8	.1
29. CPI Ac	1.8	1.4	1.4	.4	1.3	1.3	.8	.9	1.4	3.4	1.4	1.0	2.9	2.4	2.2	2.3	.1
30. CPI Fx	1.0	2.2	1.2	.4	.4	.4	.2	.4	1.3	1.0	1.4	.7	1.1	2.0	1.3	4.4	.1
31. CPI Ie	3.0	3.1	2.4	1.1	2.4	2.3	1.9	1.6	1.4	1.7	1.3	1.3	1.9	.7	4.3	1.7	.1
32. CPI Ai	1.8	2.2	1.7	.7	1.5	1.6	1.3	.9	1.4	1.7	.8	1.2	1.2	.8	2.9	2.0	.1
33. MMPI L	.8	1.7	.6	.2	.2	.1	.1	.6	1.4	.8	1.4	1.6	.2	.7	1.1	.7	.1
34. MMPI F	.6	1.4	.8	0	.8	.3	.2	.7	1.5	1.6	2.5	.8	.5	.7	.3	.1	.1
35. MMPI K	.2	1.3	.5	.1	.2	.1	.1	.8	1.9	1.6	1.6	1.1	.1	.8	.9	.9	.4
36. MMPI 1	.8	1.8	.4	.1	.2	.1	.1	1.4	2.2	1.7	2.4	1.4	.5	1.0	.3	.3	.1
37. MMPI 2	.4	1.9	.7	.1	.2	.1	.1	.7	3.0	1.5	2.6	2.3	.6	1.1	.2	.9	.1
38. MMPI 3	.4	.9	0	0	.1	0	.2	.4	1.9	1.8	3.5	.9	.2	2.0	.4	.4	0
39. MMPI 4	.3	1.6	.8	0	.1	0	0	.4	2.7	3.3	1.3	1.6	.4	1.5	.9	.3	0

[a]Full Test Names Follow This Table

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 17 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Interviewing & Counseling Skill	Oral Communication	Facility with Another Language	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Personal Organization	Tolerance of the Job of Negative Aspects	Intelligence & Judgment	Flexibility	Physical Abilities
40. MMPI 5	.6	1.7	.3	.2	.3	.2	.2	1.3	1.5	.8	.5	1.4	.2	.4	.3	.7	0
41. MMPI 6	.6	1.4	.2	0	.1	0	0	.8	1.1	.3	.9	.9	0	.7	.1	0	0
42. MMPI 7	.8	2.5	1.3	.1	.2	.7	.1	1.5	1.7	2.1	2.7	2.8	3.9	2.0	1.2	1.8	0
43. MMPI 8	.6	2.2	.9	0	.1	0	0	.9	2.4	1.5	2.0	1.8	.8	1.9	1.4	.2	0
44. MMPI 9	.5	1.7	.7	.2	.4	.3	.3	.7	2.0	2.1	2.2	2.2	1.1	.7	.8	.7	0
45. MMPI 0	.2	2.3	1.9	.1	.1	.1	.1	1.1	4.6	1.4	1.7	3.0	.2	.2	.4	1.1	0
46. MMPI 0-H	.7	.8	.2	.2	0	0	0	.2	2.1	.9	2.8	.9	1.0	1.3	0	1.2	0
47. GPI Coop.	.4	2.8	.7	.2	.1	.1	.1	.8	4.4	1.5	2.2	1.4	0	1.1	0	.9	0
48. GPI Conscient	.1	.5	0	0	0	.4	.2	.6	.4	2.3	1.4	1.2	.3	1.0	.8	1.4	0
49. GZTS Gen. Act	.4	.8	.8	.2	.2	.3	.2	0	1.6	1.5	.7	1.9	.9	1.0	.2	.3	.2
50. GZTS Ascend.	.4	2.5	2.4	.2	.2	.2	.2	.4	3.3	1.4	1.4	5.4	.4	.5	.5	.2	0
51. GZTS Emot. Stab	.6	1.8	1.4	.1	.1	.1	.1	.5	2.4	2.1	3.9	1.6	.8	2.3	.6	1.4	0
52. GZTS Objectiv.	.6	2.4	.9	.1	.2	.2	.2	.8	2.1	1.3	4.1	2.4	.2	2.2	.5	1.6	0

[a]Full Test Names Follow This Table

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 17 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Interviewing & Counseling Skill	Oral Communication	Facility with Another Language	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Personal Organization	Tolerance of the Job of Negative Aspects	Intelligence & Judgment	Flexibility	Physical Abilities
53. GZTS Pers. Rel.	.4	2.1	.4	.2	.1	.1	.1	.4	3.6	1.5	1.2	1.4	.3	1.5	.1	1.2	0
54. GZTS Restraint	.4	.9	.5	.1	.1	.2	.2	.7	1.1	3.5	2.2	1.0	2.7	1.3	1.3	.6	0
55. DPQ Unfrien	.1	1.7	.1	0	0	0	0	.8	3.2	1.4	2.0	1.0	.2	1.8	.2	1.0	0
56. DPQ Authorit.	.4	1.4	.1	.2	0	0	0	.6	1.4	2.9	1.1	.7	1.1	2.3	.3	1.4	0
57. DPQ Stress R.	.6	2.1	.8	.1	.1	.1	.1	.4	1.4	1.8	5.3	1.9	.5	2.8	.4	1.6	0
58. DPQ Aggression	.2	2.3	.2	0	0	0	.5	.2	3.4	2.2	2.3	1.6	.4	1.3	0	1.4	0
59. DPQ Impulsive	.5	1.0	.8	.1	0	0	0	.2	1.7	3.0	1.5	.5	4.0	1.6	1.2	2.2	0
60. DPQ Control	.3	.6	.1	0	0	0	.2	.2	1.1	3.2	1.5	.4	4.7	1.5	1.5	1.6	.2
61. DPQ Tradition	.3	1.4	.1	.1	0	0	0	.2	1.2	3.6	.4	1.3	1.2	3.0	.4	2.2	0
62. APQ	.1	1.6	1.3	0	0	0	0	.4	2.9	.3	4.9	1.4	.3	2.7	.2	1.1	0
63. BQ Soc. Inr	.4	2.4	1.8	.1	0	0	0	.4	4.0	.2	1.2	2.8	.1	.4	0	.5	0
64. BQ Agg/Indep	.1	1.9	1.7	0	0	.2	0	.4	2.7	.3	.5	4.0	.2	1.1	.8	1.0	0
65. BQ Soc. Dep.	.2	1.7	.4	0	0	.2	0	.4	1.6	1.0	1.3	2.3	.2	1.7	.3	.3	0

[a]Full Test Names Follow This Table

Table 3.5 Average Usefulness Ratings of 72 Tests as Measures of 17 Before Hire Worker Characteristics

Test/Scale[a]	Knowledge of Psychology	Interviewing & Counseling Skill	Oral Communication	Facility with Another Language	Reading	Written Communication	Quantitative Skill	Skill in Observing Behavior	Social Skills	Motivation/Dependability	Stress Tolerance	Dominance/Self-Esteem	Personal Organization	Tolerance of the Job of Negative Aspects	Intelligence & Judgment	Flexibility	Physical Abilities
66. BQ Soc. Lead.	.1	1.9	1.8	.1	.2	.1	.2	.3	4.4	.9	1.2	3.9	.2	.5	.4	.2	0
67. BQ Adjust	.5	1.4	.6	0	0	0	.2	.4	2.5	1.2	2.5	2.9	.2	1.2	.3	.4	.1
68. BQ Exp. Neg.	.4	1.4	.5	0	0	.1	0	.1	2.4	.5	1.4	.8	0	1.2	.2	.1	0
69. Self-Esteem	.6	1.5	1.0	.1	.3	.3	.3	.2	1.9	.7	1.8	3.5	.4	.9	1.3	.4	.2
70. Empathy	1.3	4.8	.8	0	0	0	0	2.4	4.5	.5	.5	1.5	.2	.9	.5	1.9	0
71. CO Interest	.7	1.3	.3	.2	0	0	0	1.3	2.3	1.8	2.1	1.1	.2	1.9	.7	1.5	0
72. Occup. Role	.8	1.0	.7	0	.1	0	.1	.1	1.7	1.2	.7	.7	.8	1.7	.2	.6	.3

[a]Full Test Names Follow This Table

Table 3.5 (continued)

Employee Aptitude Survey

1. Verbal comprehension
2. Numerical ability
3. Visual speed and accuracy
4. Space visualization
5. Verbal reasoning
6. Symbolic Reasoning

Differential Aptitude

7. Abstract reasoning
8. Language usage
9. Verbal reasoning (analogies)

Edison Electric Institute

10. Following directions
11. Tables and graphs

Educational Testing Service

12. Hidden figures
13. Card rotations
14. Map memory
15. Building memory
16. Following directions

17. Watson-Glaser Critical Thinking Appraisal

18. Correctional Officer Reading Comprehension Test

19. Otis-Lennon Mental Ability Test

20. Wesman Personnel Classification Test, Part 1

21. Comprehensive Ability Battery Rote Memory

Table 3.5 (continued)

California Psychological Inventory

22. Dominance
23. Sociability
24. Well-being
25. Responsibility
26. Socialization
27. Self control
28. Tolerance
29. Achievement via conformance
30. Flexibility
31. Intellectual efficiency
32. Achievement via independence

Minnesota Multiphasic Personality Inventory

33. Lie scale
34. Infrequency
35. Validity
36. Social introversion
37. Hypomania
38. Schizophrenia
39. Psychasthenia
40. Paranoia
41. Masculinity-femininity
42. Psychopathic deviate
43. Conversion hysteria
44. Depression
45. Scale 0
46. Scale 0-H

Gordon Personal Inventory

47. Cooperativeness
48. Conscientiousness

Guildford-Zimmermann Temperament Survey

49. General activity
50. Ascendancy
51. Emotional stability
52. Objectivity
53. Personal relations
54. Restraint

Table 3.5 (continued)

Multidimensional Personality Questionnaire

55. Unfriendly world
56. Authoritarianism
57. Stress Reaction
58. Aggression
59. Impulsiveness
60. Control
61. Traditionalism

62. Activity Preference Questionnaire

Owens Biographical Questionnaire (unpublished)

63. Social introversion
64. Aggressiveness/independence
65. Social desirability
66. Social leadership
67. Adjustment
68. Expressing negative emotion

69. PDRI Self Esteem

70. Hogan's Empathy

71. Correctional Officers' Interest Blank

72. PDRI Occupational Role Preferences.

To summarize the results of the test to worker characteristic ratings, the 8-point scale was reduced to four categories. Tests whose average usefulness rating fell between 0.0 and 0.5 were said to have no relationship to the worker characteristic, tests with ratings between 0.5 and 2.5 were said to have a low relationship, between 2.5 and 5.5 a moderate relationship, and tests with an average rating of 5.5 or above were said to be highly related to the worker characteristic. Table 3.6 presents the number of tests having each degree of relationship to each of the 17 before-hire worker characteristics. Most worker characteristics were judged moderately related to several tests. There was one exception to this finding: Physical Abilities, which was essentially unrelated to any of the 72 tests.

Linking Tests to Task Categories

Once the task - category to worker characteristic and worker characteristic to test ratings were made, the stage was set to link tests to task categories through the intervening worker characteristics. This linking of tests to task categories required the simultaneous examination of both sets of ratings. The strategy adopted for linking tests to tasks was to set criteria that were as stringent as possible, while at the same time, ensuring that at least one test was linked to each task category. Procedurally, each task category was considered in turn, and all of the worker characteristics whose average ratings met or exceeded the specified criterion (see Table 3.7) were considered linked to that task category. Each of the worker characteristics so linked were themselves examined and all of the tests whose average usefulness rating for that worker characteristic met or exceeded the second criterion were linked to the task via that worker characteristic. For example, for the first task category (Reviewing charges, previous record), one worker characteristic (Reading) had an average importance rating of 4.0 or higher. For Reading one test (CO Reading) received an average usefulness rating of 5.5 or higher.

Once all test to task category linkages were made, the list of linked tests was revised to eliminate duplicates (due to worker characteristics being linked to more than one task), yielding the final set of tests (see Table 3.8) that comprise the experimental selection battery for probation officers.

Table 3.6

Reliabilities and Ratings Summary of Test to Worker Characteristic Ratings

Before-hire Worker Characteristic	Strength of Task Relationships[a] None	Low	Med.	High	Rating rel.
A. Knowledge of psychology	28	36	8	0	.88
B. Interviewing & counseling skill	8	52	12	0	.82
C. Oral communication	22	43	7	0	.89
D. Facility with other Languages	54	17	1	0	.83
E. Reading	49	15	7	1	.97
F. Written communication	51	13	7	1	.96
G. Quantitative skill	49	21	1	1	.95
H. Skill in observing behavior	23	47	2	0	.72
I. Social skills	15	40	17	0	.91
J. Motivation/dependability	25	36	10	1	.90
K. Stress tolerance	22	38	12	0	.89
L. Dominance/self-esteem	19	42	10	1	.90
M. Personal Organization	34	33	5	0	.89
N. Tolerance of negative job aspects	21	45	6	0	.78
O. Intelligence & judgment	24	35	8	5	.95
P. Flexibility	22	48	2	0	.74
Q. Physical abilities	72	0	0	0	.08

[a]See text (pg. 31) for full descriptions.

Table 3.7

Linkage Criteria and Summary for Test to Worker Characteristic and Worker Characteristic to Task Category Ratings

Task Category[b]	Task Category to Worker Characteristic		Worker Characteristic To Test	
	Criterion	N Linked	Criterion	N Linked
1	4.0	1	5.5	1
2	4.0	2	5.5	0[a]
3	4.0	2	5.5	7
4	4.0	4	5.5	5
5	4.0	2	5.5	6
6	4.0	2	5.5	0[a]
7	4.0	2	5.5	6
8	4.0	2	5.5	6
9	4.0	2	5.5	5
10	4.0	5	5.5	5
11	4.0	0[a]	5.5	0[a]
12	4.0	4	5.5	5
13	4.0	0[a]	5.5	0
14	4.0	0[a]	5.5	0
15	4.0	4	5.5	5
16	4.0	3	5.5	3
17	4.0	1	5.5	2
18	4.0	2	5.5	5
19	4.0	2	5.5	7
20	4.0	0[a]	5.5	0
21	4.0	2	5.5	5
22	4.0	1	5.5	5
23	4.0	1	5.5	5
24	4.0	1	5.5	0[a]
25	4.0	4	5.5	5
26	4.0	7	5.5	5
27	4.0	3	5.5	7
28	4.0	2	5.5	2
29	4.0	2	5.5	1
30	4.0	1	5.5	1
31	4.0	2	5.5	5
32	4.0	3	5.5	7
33	4.0	3	5.5	5
34	4.0	1	5.5	0[a]
35	4.0	1	5.5	0[a]
36	4.0	0[a]	5.5	0
37	4.0	5	5.5	6

[a]Relaxing the linkage criteria for these KSAs & tasks would not have led to the inclusion of tests otherwise not included in the final selection battery.

[b]See Table 3.2 for Task Category labels.

TABLE 3.8

Cognitive and Noncognitive Marker Tests for the Experimental Selection Battery

Marker	Test Description
Cognitive Ability	
EAS Verbal Comprehension	A 30-item vocabulary test with a five-minute time limit designed to measure the ability to use words in oral and written communications.
EAS Numerical Ability	A 75-item arithmetic test in three parts with a total test 10-minute time limit. The test measures a person's facility with the arithmetic operations of addition, subtraction, multiplication, and division using problems combining integers, decimals, fractions, and percentages.
DAT Language Usage	A 50-item test in which persons must identify which of four sections of a sentence contains an error in punctuation, grammar, or capitalization. The time limit for this test is 25 minutes.
Watson-Glaser	A 50-minute test that measures ability in five areas: inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments.
Otis-Lennon	An 80-item multiple choice test that measures general intellectual prowess. Items consist of verbal and figured analogies, math word problems, vocabulary problems, etc.

Table 3.8 (continued)

Marker	Test Description
Wesman Verbal	A 20-item analogy test that measures the ability to reason through analogy and the perception of relationships.
DAT Verbal Reasoning	A 50-item analogy test. Persons must identify which of five pairs of words completes a sentence and makes an analogy true.
EAS Verbal Reasoning	Six problems, each with five items. Each problem consists of a series of six facts followed by five possible conclusions. Persons are to indicate which conclusions are true, which are false, and which are indeterminate.
CO Reading	A 50-item Reading test consisting of several short passages, each followed by factual questions. Measures the ability to read and recall what was read.
Noncognitive Tests	
CPI Dominance	A 46-item subscale of the California Psychological Inventory that identifies individuals who behave in a dominant ascendant manner and who, in interpersonal situations, can take initiative and exercise leadership in a forceful, self-confident manner.
CPI Sociability	A 36-item subscale of the California Psychological Inventory that indicates a person's tendencies to be outgoing and participative and to seek out and enjoy social contact without being shallow or superficial.
CPI Responsibility	A 42-item subscale of the CPI measuring a person's tendency to believe that life is best governed by reason and to behave in a responsible, self-disciplined manner.

Predictor Development

As was detailed above, PDI and PDRI staff psychologists rated the usefulness of 72 tests of cognitive or noncognitive abilities as measures of the worker characteristics necessary to successfully perform the probation officer job. Analysis of the ratings showed them to be very reliable and indicated that 12 of the 72 tests were likely to be useful predictors of job performance. Three of the tests were measures of noncognitive abilities. The remaining 9 were measures of cognitive abilities. The contract between the Board of Corrections and PDI allowed for the use of existing noncognitive tests in the final selection battery, but required all cognitive ability tests to be newly developed so that the copyright to the tests could be assigned to the California Board of Corrections. In accordance with the contract, the three noncognitive measures were permissible for use in the selection battery "as is", contingent only upon appropriate permissions and royalty agreements being negotiated with the test publishers. The remaining 9 tests, being measures of cognitive ability, could not be used in the final battery, and so required the development of new tests designed to measure the same abilities.

Development of the new cognitive ability tests involved three steps. First, items were drafted and preliminary test booklets assembled. Second, the newly constructed tests were pretested in a sample of college students (administered with the marker tests they were designed to replace). Third, data from the pretest were used to refine the tests into a form ready for use in the experimental selection battery. Each of the three steps is detailed below.

Test Construction. In order that the new tests parallel as closely as possible the format and content of the marker tests they were to replace, the first step in test construction was to outline a test development plan for each test. The development plans (see sample in Figure 3.2) described the ability to be measured, the number and type of items to be included, whether the test was intended to be a speed or power test, the proposed time limit for the test, and any other comments that might be useful in guiding construction of the test. Once the test development plans were completed, test construction proceeded similarly for each of the 7 tests. Only 7 tests were developed because the abilities measured by the 9 identified marker tests overlapped a great deal.

Using the test construction plans as blueprints, PDI staff members drafted 2 to 2 1/2 times the number of items needed for the final tests. The surplus of items ensured that the tests could be significantly revised and still contain sufficient items to be reliable and valid.

Figure 3.2

Test Development Plan for Numerical Ability Test

Ability Measured
Mathematical knowledge and computational ability.

This test is to measure knowledge of the four basic mathematical operations (addition, subtraction, multiplication, and division). Applicants will be required to apply this knowledge to simple problems involving whole numbers, fractions, decimals, and percentages.

Marker Test
Employee Aptitude Survey Test 1.

Test Format
Test should consist of 50 multiple choice items, each with four distractors. This is intended to be a speeded test with a time limit, excluding instructions, of five minutes.

Comments
This test is intended to roughly parallel the EAS marker but with fewer items and a slightly different item mix. Distractors should be devised to be the results that would obtain from common errors. For example, a problem such as 1/3 x 1/2 might have as distractors 1/5, 2/5, 2/6, and the correct answer, 1/6. In this way, the test will tap both the knowledge (what to do) and computational (actually doing it) aspects of numerical ability.

Once items were drafted, two preliminary forms of each test were assembled and scoring keys developed and checked. Item distractors were arranged so that the correct response occupied each possible position (i.e., A, B, C, D, or E) with approximately equal frequency. Each test was then reviewed by PDI staff members for obviously biased or unfair items and the scoring keys rechecked for accuracy. The resulting preliminary tests were printed in sufficient quantity for pretesting. Preliminary test construction yielded two forms of each of the 7 cognitive ability tests to be used in the experimental battery.

Pretesting the Preliminary Tests. Arrangements were made with instructors at four-year and community colleges to permit their students to serve as participants in the pretest sessions. In return the students received extra credit toward their class grades. Each preliminary test was administered, under the same conditions to prevail during actual test use, to groups of 15-30 students. No student completed more than two of the preliminary tests. Table 3.9 presents the number of students completing each of the preliminary tests. A single version of the experimental selection battery was prepared for use with all three jobs. This was considered necessary as validation testing for the three jobs occurred simultaneously, and it was critical to prevent any accidental cross administration of, say, the corrections officer tests to probation officers.

During pretesting, students were administered the appropriate marker test (under the time limits and instructions provided by the publisher), followed by both forms of the preliminary test. For the preliminary tests, students were instructed to indicate the item they were working on at the end of the time limit, but were then permitted to continue working until they had completed all items. This procedure allowed the time limits to be gauged as well as permitting the collection of data about every test item. Because these data were used to compute the items statistics that were in turn used to select the final items, it was important to ensure that psychometrically sound items were not eliminated from the final tests simply because they appeared near the end of a preliminary test. Such elimination might have occurred had the pretesting time limits been so short as to prevent students from attempting all of the items.

Item Analysis and Revision. Using the data collected in the pretest sessions, three statistics were computed for every preliminary test item. The difficulty of each item (proportion of persons passing the item) was computed as was the correlation of each item with the total score. Item-total correlations were computed as the point-biserial correlation between an item (scored dichotomously as pass or fail) and the score received on the sum of the remaining items. Finally, the frequency with which each distractor was chosen was determined.

Table 3.9

Pretest Results and Item Statistics for the Experimental Selection Battery

Test	Number of Students Tested	Avg. Item Diff.	Item-Test Corr.	N Items	rel.[a]	Validity[b]
Observational Memory	30	.90	.10	30	.44	.31
Numerical Ability	19	.74	.57	35	.92	.75
Vocabulary	21	.66	.37	35	.83	.82
Language Usage	29	.78	.51	35	.90	.88
Reading Comprehension	30	c	c	70	c	c
Analogies	16	.78	.62	35	.93	.69
Verbal Reasoning	25	.57	.26	35	.80	.72

[a]Internal consistency via KR-20.

[b]Correlation with marker test.

[c]Data unavailable due to method of revision -- see text (p. 42) for details.

The item statistics were used to select items to appear in the experimental selection battery. Prior to this step, the average number of items completed in the proposed time limit was computed for each test. These results showed that students were working the items quickly enough to permit a switch from speed to power measures. Speed measures are those where the constituent items are so easy that, given sufficient time, everyone should be able to get every item correct. For speed tests, time limits are set that are much shorter than is necessary for most people to complete all test items. Speed tests are predicated upon the idea that people with more or higher levels of the ability being measured will be able to work the items faster. Thus, a higher score on a speed test is an indication of better ability. On the other hand, power tests have time limits that permit most persons to complete all items. On power tests the items vary in difficulty, with some being so easy that nearly everyone can get them correct while others are much more difficult.

To summarize, speed tests depend upon time limits to distinguish among persons of varying ability while power tests depend upon item difficulty. The advantage of power tests, especially in a project such as this, lies in their independence from time limits.

In order for any test to give an accurate measure of ability, or in this case the potential to be a successful probation officer, the test must be administered under as nearly identical conditions as possible every time it is used. With speed tests, especially those with relatively short time limits, it is very difficult to ensure that the time limits are rigidly followed during every administration. Because they have no strict time limits, power tests are more desirable for projects such as this. As a result, each test in the experimental selection battery (with two exceptions discussed below) contained approximately the average number of items completed in the pretesting time limit. The total time allowed for the resulting battery was then doubled. Thus, the final time limits were set so that virtually every person would be able to complete all items within the allotted testing time. In this manner, the tests were transformed from highly speeded to power tests, with an accompanying reduction in administrative complexity.

Once the desired number of items was determined for each test, the pretest data were used to screen items and to reject those that were not of appropriate psychometric quality. Several factors were considered in the course of this screening. For example, items for which the keyed alternative correlated negatively with total scores, or where the item-total correlation for the keyed alternative was less than that for the distractor, were rejected. No initial screening was done on the basis of item difficulty. The items remaining after screening formed the final pool of psychometrically sound items from which item selection was made.

In order to ensure an overall difficulty level appropriate for the intended job applicant population, items were selected based upon their item difficulties. For example, five items with difficulties between 1.0 and .90, five with difficulties between .90 and .80, and so on, were chosen for each test. Within difficulty levels, items with the highest item-total correlations were selected.

Once the desired number of items had been selected, the reliability (KR-20) and validity (correlation with the published marker test) of each test were determined. If these were satisfactory, test construction was complete. If not, the correlation of each item with the marker test total score was computed (for all items in the final item pool) and this index used as the basis for item selection. That is, within a specific range of item difficulty, the items which correlated most highly with the marker total score were selected. Table 3.9 summarizes the psychometric properties of the tests that make up the experimental selection battery.

As was true for the preliminary tests, items in the final tests were arranged so that no patterns of correct answers were present and so that the correct alternative appeared in each possible position with approximately equal frequency. Finally, items were arranged (approximately) in ascending order of difficulty.

Concern that the tests might prove too difficult for the applicant population led to the inclusion of five addition items in each test. All of the additional items had item difficulties of 1.0 in the pretest sample. These items were included to ensure that each test had sufficient breadth to prevent restricting the range of test scores due to floor effects. "Floor effects" refers to a situation in which the items of a test are all too difficult for the test taking population. If this occurs, scores on the test will all be very low and there will be a large number of persons who receive the same score. A large number of tied scores reduces the variance of the scores and puts a limit on the maximum possible validity of a test. In order not to limit the potential validity of the tests, several very easy items were added to each test.

Special Revision Methods. One of two tests not revised as described above was the Observational Memory Test. Rather than selecting a subset of the 30 items in the preliminary version of this test, it was decided to retain the test intact. Data on the pretest time limits showed that the test was easily completed in a minimal amount of time, therefore little was to be gained by shortening the test. The time limit results, coupled with item statistics indicating that the items were too easy, led to a reduction in the suggested time limits. Rather than having two minutes to study the stimulus figures to be memorized, only one minute was allowed. The amount of time allowed to complete the ten items in each test section was also reduced from three to two minutes. In addition to the changed time limits, the items were made more difficult by increasing the complexity of the stimuli to be

memorized and by requiring more subtle discriminations between correct and incorrect alternatives. No other revision of the test was undertaken, and it remained the only test to depend upon speed rather than power to distinguish among incumbents.

The other test for which a variation in the revision procedures was made was the Reading Comprehension test. Pretest data for the reading marker test (CO Reading) revealed that the test was too easy for the pretest sample, leading to very little variation in test scores. Much the same was true for two of the three parts of the preliminary reading test. As a result, the item statistics were judged too unreliable to serve as a guide to test revision. One section of the preliminary test (Part 2) in particular was clearly not working as intended. This part of the test required subjects to make inferences and draw conclusions from minimal amounts of information. The problem with this part of the test was that no clearly justifiable key could be agreed upon and subjects were about equally divided in their choice of distractors.

The experimental selection battery Reading Comprehension test was made up of the six passages that PDI staff judged to be most appropriate for the incumbents to be tested. The 10 to 12 items following each passage were an approximately 75%-25% mix of "search and find", and reasoning items. The former required the recall of answers to factual questions, while the latter required simple inferences and conclusions to be drawn from the material presented in the passages. These reasoning items are the only remaining vestige of Part 2 of the preliminary scale. Enough items were included in each passage to allow the test to be further refined (based on item statistics and using the procedures described above) using data collected during the validation phase of the project.

Final Test Battery

The newly developed cognitive ability tests were assembled in three booklets. The first booklet contained the Observational Memory test. This test required strict adherence to time limits and so was kept separate from the other tests. The second booklet contained the Reading Comprehension test. A time limit of 30 minutes was set for this booklet. On the basis of the pretest data, 30 minutes was judged to be sufficient to allow virtually all incumbents to complete all test items. The test consisted of six passages, each with 10 to 14 items. Other than the 30-minute limit for the entire booklet, incumbents were not constrained in the amount of time they could spend on any one passage.

The third test booklet contained the Numerical Ability, Language Usage, Vocabulary, Analogies, and Verbal Reasoning Tests. Because testing was to occur nearly simultaneously for three jobs, it was feared that separate test batteries might be misadministered. To prevent such an event from occurring, a single battery containing the union of tests developed for

all jobs was assembled and administered to probation officers as well as to juvenile counselors and corrections officers.

Each test in Booklet 3 contained 35 items (175 items in all). A time limit of 60 minutes was set for the entire booklet. This allowed ten minutes per test (twice the estimated time necessary to complete 35 items) and left an additional ten minutes of testing time. Incumbents were free to divide the 60 minutes among the five tests as they saw fit. However, there were prominent warnings that no more than ten minutes should be spent working items for any given test until all tests had been attempted.

Noncognitive Tests

Noncognitive marker scales likely to predict job performance were identified at the same time and using the same methods as described for the cognitive marker tests. Three of the five tests thus identified are shown in Table 3.8. The publisher of each test was contacted to secure permission to use their test in this project and to ascertain whether the Board of Corrections would be granted permission to score the tests at the agency level. For two of the marker tests (Guilford-Zimmermann Temperament Survey and Hogan's Empathy), it seemed unlikely that the necessary permissions would be obtained. The Guilford-Zimmerman Temperament Survey was omitted from the experimental selection battery, but the California Psychological Inventory items that are part of Hogan's Empathy Scale (some of which overlap with other CPI scales in the battery) were retained.

Although not identified as a potentially useful scale in the analysis conducted to identify noncognitive marker tests, the Correctional Officers' Interest Blank (COIB) was selected for inclusion in the experimental selection battery based on its proven performance in similar studies. The Socialization scale of the California Psychological Inventory and the Control and Stress Reaction Scales of the Multidimensional Personality Questionnaire were also included in the experimental selection battery. These tests were targeted to the juvenile counselor or corrections officer positions but were also administered to probation officers.

With the exception of 18 items from the COIB, all of the items from the noncognitive tests were in true/false format. These items were edited to change all occurrences of the pronoun "I" to "you". These changes necessitated other minor editorial changes so that the items read well. A few items were omitted because of their invasive nature. These items were potentially self-incriminating (e.g., "You have never been in trouble with the law") or dealt with topics too personal for use in an employment setting (e.g., "You have been in trouble one or more times because of your sex behavior").

Editing was kept to a minimum. STC staff members also reviewed the items for biased or objectionable items, leading to the deletion of five items, as well as several minor editorial changes in item content.

All noncognitive items were assembled into one booklet with two sections. The first section contained the 228 true/false items from the noncognitive tests. Items from the different scales were mixed together unsystematically. The second section contained 18 items from the COIB. These items required incumbents to pick from among three alternatives the activity they would like most and the activity they would like least. Based on an estimated response rate of five items per minute (12 seconds per item), a time limit of one hour was set for the noncognitive items booklet.

Sample Items Booklet

To provide the incumbents in the validation sample an opportunity to become familiar with the cognitive ability items in the experimental selection battery, a booklet with illustrative items from each cognitive ability test was assembled. For each test the purpose of the examination was briefly explained, two or three sample items given with the correct answers explained, and eight to ten practice items (with answers) given. This booklet was distributed to all incumbents tested and was the prototype of a similar booklet to be made available to future job applicants as a preview of the selection battery.

CHAPTER FOUR: CRITERION DEVELOPMENT

A criterion-related validation strategy was used in this study. In a criterion-related approach, tests are valid if test scores can be demonstrated to be statistically related to a reliable measure of job performance. The agencies participating in this study used different methods of assessing the performance of their incumbents. Thus, it was not possible to use existing performance ratings as the criterion for judging the validity of the experimental tests. Even if a standard method of assessing incumbent performance had been available, it would not have proved useful because professional standards mandate that criterion-related studies include a relevant and reliable criterion measure. Existing job performance data were of unknown relevance and unknown reliability. Thus, it was necessary to develop new measures of job performance to be used as the criterion for assessing the validity of the predictors.

To ensure the relevance of the criterion measures, the job analysis data were used to guide criterion development. These data provided the clearest picture of the most important aspects of the probation officer job and thus provided an ideal starting point for criterion development. The goal of criterion development was to create a series of scales that fully cover the relevant aspects of the probation officer job. These scales would then be used by supervisors to rate the job performance of the incumbents who completed the experimental selection battery.

A critical incidents approach was taken to criterion development. In this approach, probation officer supervisors and incumbents wrote behavioral incidents representing examples of poor, average, and excellent performance for critical aspects of the probation officer job. These incidents, when analyzed and summarized, provided the benchmarks that were used by other supervisors to rate the performance of incumbents in these important areas. Details of the critical incident methodology used to develop the performance criterion measures are given below.

Collecting Critical Incidents

Supervisors and incumbents were recruited to participate in this phase of the study in such a way that large and small agencies from both northern and southern California were represented. Participating incumbents and supervisors were intimately familiar with the duties of probation officers and were articulate and able to express themselves well in writing. Twenty-five incumbents and 30 supervisors attended 1/2-day workshops conducted by a PDI psychologist. During the workshops, the critical incident process was explained and the critical incidents written.

To guide their thinking about the job, participants were provided a summary of the job analysis consisting of descriptions of the task categories developed for use in linking worker characteristics to task categories (See Chapter Three).

The task categories were explained to the participants who were then provided forms (See Figure 4.1) on which to record their critical incidents. These forms gave general instructions about writing critical incidents and provided space to: 1) describe the circumstances leading up to each incident; 2) describe what the incumbent did during the incident; 3) rate the performance on a 1 (extremely ineffective) to 9 (extremely effective) point scale; and 4) indicate into which (if any) of 14 general task categories the incident fit.

Each supervisor and incumbent was able to generate approximately 11 incidents yielding a total of 614 critical incidents from the 55 supervisors and incumbents. These incidents provided the raw material from which development of the criterion measures proceeded.

Revising and Retranslating the Critical Incidents

The incidents were edited to emphasize the behaviors involved, to separate double incidents, and to ensure that all incidents were described in a similar format. PDI staff members rationally grouped the incidents into categories based upon the behaviors involved in each incident. The goal was to create categories which included only those incidents revealing varying levels of effective performance for similar or related job tasks. Fourteen such categories of incidents were created and are described in Table 4.1. To corroborate these job performance categories and to ensure that each incident was correctly assessed as to its degree of effectiveness, the 614 incidents were retranslated by probation officer supervisors. During retranslation a second set of supervisors were given a list of the edited incidents and descriptions of the categories developed from the incidents. The supervisors were asked to place each incident in the most appropriate category. Retranslation was necessary to be sure that, from the perspective of persons familiar with the job, each incident was correctly grouped and to ascertain the degree of effectiveness each incident represented. For reasons of administrative feasibility, the 614 incidents were assembled into two booklets, each with approximately 350 incidents (the booklets contained overlapping incidents to permit the data collected from the two booklets to be equated). The booklets contained detailed instructions for the retranslation task and provided descriptions of the 14 performance categories to be used in making the ratings. Each booklet contained incidents believed to belong in each of the 14 categories.

Figure 4.1

PERFORMANCE INCIDENT FORM

Thinking back, focus your attention on any specific thing that you saw a person do that made you think of him/her as a highly effective worker, an average worker, a very ineffective worker, or any shade of effectiveness in between. In other words, think of a specific example of job behavior that gave you some feeling for how well the individual was doing his/her job. (Please do not record the names of the people involved.)

1. What were the circumstances leading up to the incident?

2. What did the individual do that made you think he/she was a good, average, or poor performer?

3. How effective do you feel this example of performance was (Circle a number.)

1	2	3	4	5	6	7	8	9
extremely ineffective		ineffective		about average		effective		extremely effective

4. What, if any, general category would the above action fit into?

Table 4.1

Probation Officer Performance Dimensions Used During Retranslation

---

A. INVESTIGATING

Thoroughness in obtaining information from files, records, employer, other agencies, family; checking with other states, FBI; verifying probationers' statements; following up on leads.

B. MAKING RECOMMENDATIONS

Making recommendations for detention/release, placement, sentencing, restitution and other fines, terms and conditions of probation, and other recommendations as required; using information from case law, Penal Code, Determinate Sentencing Manual, and other relevant sources; taking appropriate information and circumstances into account.

C. WRITING REPORTS

Writing or dictating clear and accurate court reports; preparing court reports efficiently and on time; clearly organizing relevant information; using good grammar and sentence structure.

D. HANDLING PAPERWORK

Accurately, promptly, and conscientiously completing case notes and other paperwork; keeping records organized and up-to-date.

E. ESTABLISHING RELATIONSHIPS WITH NEW PROBATIONERS

Clearly communicating conditions of probation; gaining probationers' cooperation and respect; skill in obtaining information; providing probationers with help in meeting conditions of probation (e.g., employment, lodging. transportation, treatment) as appropriate; setting up schedule for payment of fines.

F. MONITORING AND ENFORCING COMPLIANCE

Maintaining regular contact with probationer; checking probationers' compliance with conditions (e.g., employment, residence, treatment); enforcing conditions of probation; making home visits; deciding when to file a violation of probation.

Table 4.1 (continued)

G. **DEALING WITH HOSTILE PEOPLE AND OTHER EMERGENCIES**

Taking appropriate action/precautions to prevent escalation and reduce danger; requesting assistance when appropriate; reducing versus increasing tension/conflict; remaining calm under stress.

H. **COUNSELING**

Dealing with troubled/upset probationers, family, victims, others; listening; defining/clarifying/identifying problem; helping to resolve problem or referring as appropriate.

I. **ALERTNESS**

Being alert to signs of problems; recognizing signs of emotional disorder, family problems, drug use, etc; noticing suspicious behavior; making connections between cases/events; anticipating and preventing potential problems; recognizing problem (e.g., non-reporting) probationers.

J. **INFORMING/ADVISING NON-PROBATIONERS**

Informing parents/guardians of minors about circumstances and conditions; notifying victims of rights/court date; responding as appropriate to requests/questions from probationers' families, employers, teachers; notifying relevant others of potential danger (e.g., past and potential victims, law enforcement, school principals, employers).

K. **INTERACTING WITH ATTORNEYS/JUDGES**

Establishing and maintaining cooperative relationships with attorneys and judges; demonstrating competence by testifying effectively, providing accurate information; conducting oneself in a professional manner; gaining and maintaining the respect of legal personnel.

Table 4.1 (continued)

L. INTERACTING WITH OTHER STAFF/AGENCIES

Establishing and maintaining cooperative relationships within the department (e.g., with co-workers, supervisors, clerical staff), with other criminal justice system personnel (e.g., law enforcement, juvenile hall staff, departments in other locations), and with referral and placement resources; providing/seeking assistance when needed or when requested to do so; keeping others informed of things they should know; being considerate of co-workers; helping/training new workers.

M. PROFESSIONAL INTEGRITY

Maintaining case confidentiality; admitting/correcting mistakes; abiding by professional ethics versus, for example, accepting bribes, falsifying expense reports, informing probationer of confidential information.

N. COMMITMENT AND INITIATIVE

Demonstrating commitment to the job and respect for a department's policies and procedures; putting in more time and effort than required versus doing just enough to get by; keeping informed via meetings, memos, professional training; going above and beyond the call of duty; doing things to improve the operation or the image of a department, (e.g., translating forms, giving community lectures/presentations).

probation officer supervisors were selected to categorize the incidents into one of the 14 possible job performance categories and to rate the degree of effectiveness each incident represented. Effectiveness ratings were made on a 9-point scale. Twenty-two supervisors completed the first booklet and 22 completed the second. These supervisors represented 13 agencies and were responsible, on average, for supervising 8.6 probation officer incumbents. They had been supervisors for an average of 9.3 years and the 42 who had also served as probation officer incumbents reported an average of 8.5 years experience.

Developing Behavioral Summary Scales

For each retranslated incident, the proportion of supervisors who placed the incident in each category was computed as was the mean effectiveness rating and the standard deviation of the ratings about that mean. These data were used to revise the categorization and to select incidents to appear as behavioral examples of poor, adequate, and excellent performance. The retranslation data suggested that several of the behavioral categories should be combined. Accordingly, the following changes were made: the "Writing Reports" and "Handling Paperwork" categories were combined, all incidents from the "Alertness" category were added to the "Monitoring and Enforcing Compliance" category, the "Professional Integrity" and "Commitment and Initiative" categories were combined to form "Commitment to Profession and Department", and the "Dealing with Hostile People and Other Emergencies" and "Counseling" categories were joined to form "Dealing with Crises, Counseling".

Incidents within each category were sorted into those indicating poor, adequate, and excellent performance and used to construct 10 rating scales. Each scale (See Figure 4.2) provided a description of the performance category to be rated along with a 7-point rating scale. Ratings of 1 or 2 were labeled as "needs improvement". Ratings of 3, 4, or 5 were labeled "fully adequate", and ratings of 6 or 7 were labeled "excellent". Three to five incidents were abstracted to provide examples of the kind of behaviors that should result in a given performance rating. Instructions to raters were developed to emphasize that the behavioral examples were not checklists, but rather provided to guide the raters' thinking about the typical kinds of behaviors and therefore the level of performance of each incumbent being rated.

Other Ratings Scales

Worker Characteristic Ratings. In addition to the 10 behavioral summary scales previously described, 21 of the worker characteristics created during predictor development (see chapter three) were transformed into 7-point rating scales. Ratings of 1 or 2 indicated "low" standing (among the lowest one-third of employees), 3, 4, or 5 indicated "average"

Figure 4.2
Sample Behavioral Summary Scale

A. INVESTIGATING

Obtaining complete, accurate information from files, records, other agencies, or other individuals, and thoroughly verifying information received from probationers and others; being alert to new leads; resourcefully and persistently digging to uncover all pertinent information.

1 2 Needs Improvement	3 4 5 Fully Adequate	6 7 Excellent
. Neglects to check standard source of information, such as prior record, or to interview defendant parent, victim or other relevant person.	. Usually checks all standard sources of information and interviews all persons obviously involved in the case.	. Always checks every potentially relevant source to get information about a case.
. Neglects to verify statements of probationers, victims, family.	. Verifies questionable statements of probationers, victims, family.	. Routinely verifies statements made by probationers, victims, or family members with other sources; detects instances when others' statements or records might be in error and verifies.
. Gives up when initial check yields no useful information or when interview is difficult.	. Generally consults supplementary sources when initial check turns up no useful information and continues to probe when intervievee is uncommunicative.	. Persists in searching for alternate (even unusual) sources of information when initial checks yield no useful information; effectively persists in asking probing questions when interview is difficult.
. Fails to pick up on new leads (for example, tends not to depart from interview outline even when new information emerges).	. Follows up on all promising new leads which arise in the course of an interview.	. Recognizes and follows up on even subtle new leads which emerge during records checks, interviews, or in discussion with other criminal justice or social service personnel.

BEST COPY AVAILABLE

(in the middle one-third), and ratings of 6 or 7 indicated a "high" standing (upper one-third) on that worker characteristic.

Overall effectiveness. Finally, an overall effectiveness scale was constructed on which supervisors were asked to rate how well an incumbent met the standards and expectations of the job compared to other individuals with similar levels of experience. Ratings were made on the same 7-point scale ("Needs Improvement" to "Excellent") used for the behavior summary scales.

The final criterion rating scales consisted of 10 scales developed from the critical incidents data, 21 scales based upon the worker characteristics, and a single, overall effectiveness scale.

CHAPTER FIVE: DATA COLLECTION

Once the experimental selection battery and criterion ratings materials had been developed and assembled, the tests needed to be administered to a sample of probation officer incumbents and job performance ratings collected from the incumbents' superiors. Following are the efforts undertaken to secure a sample of probation officer incumbents of sufficient size and composition to permit an accurate and efficient assessment of the validity of the experimental selection battery.

Predictor Data Collection

Predictor data collection began with the setting of targeted sample sizes for each of the participating agencies. Sample sizes were based upon relative agency size; however, smaller agencies were requested to test a proportionally larger number of their incumbents than were larger agencies. The initial sampling plan called for testing 477 probation officers from 21 probation departments.

Department coordinators were contacted by phone to secure permission to test the requested number of probation officers. As agency participation was voluntary, a certain amount of negotiation was involved in arriving at an agreed-upon number of incumbents to be tested in each agency. In spite of the heavy and repeated requests for department resources imposed by this project, department coordinators were universally willing to go the "extra mile" to ensure that testing ran smoothly.

Information about the testing process was sent to each coordinator outlining the purpose of testing and reviewing the procedures to be involved. An important part of the information sent to the coordinators concerned the method by which incumbents were to be selected for testing. Because job performance ratings were to be collected for all incumbents tested, a selection method was devised that attempted to minimize the burden put upon each agency's supervisory staff. This sampling method called for the ordering of probation officer supervisors based upon their experience. Incumbents were then chosen by selecting all of the incumbents of the most experienced supervisor, all of the incumbents of the second most experienced supervisor, and so on until the targeted number of incumbents had been selected.

Testing was scheduled to occur during June 1984 but various organizational problems led to a number of delays so that testing was not completed until later that year. The first problem to occur was the withdrawal of two participating agencies, thereby reducing the targeted sample by 41 to 436. In order to offset this loss, PDI contacted other participating agencies in an attempt to secure additional participants. Three agencies were able to provide an additional 12 probation officers raising the targeted sample to 448. The additional step of contacting agencies that had participated during the job analysis phase but that were not involved in testing was undertaken. No additional incumbents were available from these agencies, however.

At that point, the maximum possible sample stood at 448. However, it was not reasonable to assume that 448 usable test batteries would be completed. A more reasonable estimate, allowing for persons who failed to appear for testing, etc., would be that approximately 85% of the batteries would be completed. Thus, the expected sample size stood at only 381. In order to secure additional subjects, STC consultants contacted departments that had not previously participated in any aspect of this project. These agencies supplied an additional 11 incumbents, raising the total targeted sample to 459.

By October of 1984, phone canvassing of the participating counties made it clear that no additional incumbents were available for testing. A count of the data collected to that point revealed a sample size of 296. In order to increase the number of incumbents tested, STC consultants contacted several agencies offering to reimburse them for the overtime payments they would incur as a result of testing additional incumbents ($75 per participant). Despite this offer, only 26 additional probation officers were tested, bring the total sample size to 322.

Criterion Data Collection

Soon after the experimental selection batteries were shipped to coordinators, job performance rating materials were also shipped in sufficient quantity to permit supervisors to rate the performance of all incumbents scheduled to be tested. These materials consisted of a performance appraisal booklet and two response booklets. The performance appraisal booklet introduced the performance rating task and provided general instructions about the rating process as well as detailed instructions for the completion of each rating scale. Each response booklet contained space to rate the performance of five incumbents. Each supervisor also received a cover memo providing more information about the ratings task (e.g., where it fit into the overall research effort), enlisting the supervisor's assistance in maintaining the confidentiality of the ratings he/she made, and directing the supervisor to the department coordinator in the event that questions about the ratings process arose.

Job performance ratings were returned by a total of 117 probation officer supervisors from 19 agencies. These raters reported having been employed as supervisors for an average of 12.2 years. Fourteen of the supervisors were female, the rest (101) were male. Fifteen supervisors were black, five were Asian, five were Hispanic, and 92 were white (one supervisor did not report his/her ethnic status). These 117 supervisors rated the performance of 246 incumbents on each of 9 behavioral summary scales, 21 worker characteristic scales, and the overall effectiveness scale representing a total of 7,872 ratings.

As predictor and criterion data arrived, checks were made to ensure that every probation officer incumbent tested had also been rated. In those cases where performance ratings did not arrive for a tested incumbent, department coordinators were contacted and asked to follow up with the appropriate supervisor to ensure that the ratings were made. Despite frequent requests to follow-up on ratings collection, it was not possible to obtain job performance ratings for every probation officer who completed the experimental selection battery.

Final Validation Sample

The criterion-related validation strategy employed in this study relies on job performance ratings serving as the yardstick against which the validity of the predictors is to be judged. The final validation sample (i.e., those probation officer incumbents whose data were analyzed to determine the validity of the predictors) consisted of those persons for whom both usable predictor and usable criterion data were collected. As was noted, 317 probation officers were tested, but only 296 had their job performance rated by their supervisor. Thus the maximum possible number of incumbents for whom both predictor and criterion data were available was 296. Unfortunately, there were other ways incumbent data were lost from the final validation sample. The missing data problems encountered in this project and the impact of those problems upon the size of the final validation sample are discussed in the following. Table 5.1 lists, by county, the number of incumbents tested and the number lost to the validation sample for the reasons enumerated. Following this discussion, the sample demographics are presented and the technical adequacy of the sample defended.

Missing Name. Incumbents' names were used to match predictor and ratings data. Incumbents who failed to put their names on their answer sheets prevented this match from being made for their data. A total of 4 cases were lost due to missing names.

Missing Ratings. A second way in which cases were lost from the validation sample was through missing ratings data. Despite repeated efforts to follow-up with department coordinators, ratings were not obtained for 13 incumbents who had completed the experimental selection battery.

Table 5.1

Composition of the Final Validation Sample

County	Number of POs Tested	Lost from Validation Sample: Predictors w/o Name	POs not Rated	Low Tenure	Not Well Known	Final N
Alameda	21	-	-	-	-	16
Conta Costa	27	-	-	-	-	18
El Dorado	0	-	-	-	-	0
Fresno	0	-	-	-	-	0
Humboldt	4	-	-	-	-	4
Kern	20	-	-	-	-	20
Los Angeles	27	-	-	-	1	26
Monterey	0	-	-	-	-	0
Nevada	4	-	-	-	-	4
Placer	7	-	-	-	-	7
Sacramento	28	-	1	4	-	23
San Bernardino	21	-	-	1	-	20
San Diego	45	-	1	-	-	44
San Francisco	18	4	5	2	-	11
San Joaquin	18	-	-	3	-	15
San Luis Obispo	6	-	-	-	-	6
San Mateo	30	-	6	-	-	24
Santa Barbara	12	-	-	-	-	12
Stanislaus	15	-	-	-	-	15
Tulare	11	-	-	-	-	11
Ventura	2	-	-	-	-	2
Yuba	6	-	-	-	-	6
			Final Validation Sample			284

Low Tenure. The validation sample was intended to consist of only experienced incumbents (i.e., those on the job for at least six months). The job performance of incumbents with six or fewer months experience would be difficult to rate as supervisors may not have had an adequate opportunity to observe these incumbents' performance under a variety of job conditions. In order to ensure that the job performance ratings of all incumbents accurately assessed their ability to perform the job, incumbents with six or fewer months tenure were not included in the final validation sample.

Inaccurate Ratings. When rating the performance of their incumbents, supervisors were also asked to indicate how well they knew the incumbent whose performance they were assessing. Incumbents whose ratings were made by supervisors who said that they did not know that incumbent "well at all" (i.e., rating of 4) were dropped from all analyses. In order to judge the validity of the predictors, an incumbent's true level of performance must be known. Obviously the ratings of an incumbent not at all well known by his/her supervisor could not be an accurate measure of the incumbent's true job performance.

Demographics of Final Validation Sample

As a result of the data losses noted, the final validation sample consisted of the 284 probation officer incumbents for whom accurate performance ratings were available. The demographic composition of the final sample is detailed in Tables 5.2 - 5.4. The sample is largely male and approximately one-fifth of the probation officers are members of protected groups (principally, blacks or Hispanics). One-half of the probation officers in the final sample have a Bachelor's degree and the remainder have one or more years of post-graduate education or a graduate degree. More than one-half of the sample have been probation officers for five or more years (only 5% have less than one year's experience on the job).

Table 5.2

Gender and Ethnic Status of Final Validation Sample[a]

Group	Number	Percentage
Gender		
Male	175	63
Female	103	37
Ethnic Status		
Black	32	12
Asian	6	2
Hispanic	18	7
White	219	79
Other	3	1

[a]Totals may vary due to missing data.

Table 5.3

Education and Tenure of Final Validation Sample[a]

Group	Number	Percentage
Education		
High School	2	1
Associate Degree	1	0
B.A./B.S. Degree	138	51
Some post-graduate	83	31
Post-graduate Degree	48	18
Tenure		
Less than 3 months	b	
3-6 months	b	
6 months to 1 year	15	05
1 to 2 years	24	09
2 to 5 years	76	27
More than 5 years	162	59

[a]Tables may vary due to missing data.

[b]POs with fewer than 6 months tenure were not included in the final validation sample.

Table 5.4

Relationship of Incumbents to Supervisors in Final Validation Sample[a]

Group	Number	Percentage
Length of Time Supervised		
Less than 3 months	9	3
3 to 6 months	16	6
6 months to 1 year	34	12
1 to 2 years	59	21
2 to 5 years	122	44
More than 5 years	37	13
How Well Known by Supervisor		
Very well	200	72
Pretty well	70	25
Not too well	8	3
Not at all well	b	

[a]Totals may vary due to missing data.

[b]Incumbents known "not at all well" were dropped from the final validation sample.

Technical Adequacy of Sample

One concern in conducting a criterion-related validation study is whether enough people can be tested to permit a technically adequate evaluation of the validation and fairness of the potential predictors. Several analyses were performed to determine the technical adequacy of the sample collected in this study (See "Report on the Technical Adequacy of the Final Validation Sample for Probation Officers" for details, available from the Board of Corrections). The major issue was whether the sample size provided sufficient statistical power to detect predictor validity and subgroup differences should they have existed. The analyses uniformly demonstrated the adequacy of the validation sample for all of the analyses performed during the course of the study. Thus, the 284 probation officer incumbents for whom predictor and job performance data were available provide a more than sufficient basis for judging the validity of those predictors as measures of probation officer job performance.

The analyses reported in the following chapters were based upon all available data. This ensured that the results of those analyses were as stable as possible. As a consequence, the sample size varied slightly from analysis to analysis. This occurred because an incumbent who was missing data essential to one analysis might have had all the data required for another. Thus, the sample size would have varied because the incumbent was excluded from the first analysis but included in the second. Exact sample sizes for the analyses are reported in the tables summarizing the analyses. Slight variations in sample size are to be expected for the reason given above.

CHAPTER SIX: ANALYSIS AND RESULTS

Many different analyses of the predictors and criteria were conducted to document their validity and fairness. The goal of all analyses was to obtain a combination of predictors that simultaneously:

1) would validly predict job performance;

2) was fair (i.e., did not unfairly impact on protected groups); and

3) was administratively practical for all agencies participating in the STC Program to implement.

The analyses conducted to meet the above goals can be divided into three categories: criterion analyses, predictor analyses, and validation and fairness analyses.

Criterion analyses included those necessary to determine: a) that the job performance ratings were reliable; b) how best to combine the 32 rating scales into a single measure of overall job performance; and c) whether there were specific aspects of job performance differentially predictable by one or more of the 15 tests.

Predictor analyses included those necessary for final revision of the tests (especially Reading and Observational Memory), to: a) document the psychometric quality of the tests; and b) describe the level of performance obtained by the probation officers in the final validation sample.

Validation and fairness analyses were those necessary to determine: a) how best to combine the tests from the experimental selection battery into a useful, valid predictor of job performance; and b) that the selected tests did not unfairly discriminate against candidates from protected groups.

While each of the above categories of analysis are conceptually distinct, there was considerable overlapping of analyses. Each category will be described separately although occasional reference to analyses from other categories will be necessary.

Criterion Analysis

The criterion analyses had as their basic goal the development of a reliable measure of overall job performance. This criterion would serve an important role in later validation analyses. An additional and equally important goal of the criterion analyses was to discover whether there were specific aspects of job performance that could be better predicted by single tests than by the Probation Officer Examination. The presence of such differentially predictable aspects of job performance would have broadened the usefulness of the validation research as it would have resulted in a "menu" of selection instruments from which agencies could pick according to their needs.

Data collection and cleaning. Incumbents' job performance was rated by their supervisors on 10 behavioral summary scales, 21 worker characteristic scales, and on a single, overall effectiveness scale (see Chapter Four). To preserve the confidentiality of the performance ratings, supervisors were provided pre-addressed envelopes for mailing their rating materials directly to PDI. No one other than PDI staff members and the supervisor making the ratings had access to the performance ratings of any individual probation officer.

When the ratings arrived at PDI, they were reviewed for completeness and assigned an identification number. Ratings were then keypunched. The keypunchers identified any ratings where there was a question as to the intended rating. PDI staff reviewed all ratings so identified and determined the correct rating value or set that value to "blank", indicating a missing value.

The mean rating value for each of the 32 individual rating scales is shown in Table 6.1. As can be seen, probation officers varied somewhat in the ratings they received on the 32 scales. For example, performance in the area of Oral Communications is more highly rated than is performance in the area of Safety/First Aid. It should be noted that comparisons of this sort are only appropriately made between ratings from similar scale types (i.e., worker characteristic to worker characteristic or behavioral summary scale to behavioral summary scale) and are valid only for average group performance. To compare an individual's performance ratings across scales, it would first be necessary to convert the ratings into standardized units. This is necessary because a rating of 4 on a scale with an average of 4.5 represents poorer relative performance than does a rating of 3 on a scale with an average of 2.5.

The overall performance of probation officers in the final validation sample ranges from "needs improvement" to "excellent". The mean rating for the overall effectiveness scale was 4.96 with a standard deviation of 1.17 indicating that 90% of all probation officers received ratings between 3 and 7 on the 7-point scale.

Table 6.1

Descriptive Statistics for 32 Job Performance Ratings Scales

Scale	Number of POs Rated	Avg.[a]	Stnd. Dev.
Behavioral Summary Scales			
Investigating	284	4.93	1.21
Making Recommendations	284	4.88	1.15
Handling Paperwork/Writing Reports	284	4.74	1.48
Est. Relationship with New Probationers	281	5.07	1.14
Monitoring and Enforcing Compliance	276	4.92	1.14
Dealing with Crises/Counseling	280	4.94	1.16
Informing/Advising Non-Probationers	282	4.73	1.04
Interacting with Attorneys/Judges	284	4.90	1.16
Interacting with Other Staff/Agencies	284	5.06	1.19
Commitment to the Profession, Dept.	284	4.85	1.29
Worker Characteristic Scales			
Department-Specific Knowledge	284	4.87	1.16
Knowledge of Specific Laws	284	4.71	1.22
Knowledge of Legal System, Terminology	284	4.93	1.13
Knowledge of Psychology	284	4.84	1.19
Knowledge of Chemical Use, Abuse	284	4.72	1.12
Interviewing and Counseling Skill	284	5.10	1.28
Safety, First Aid	282	4.17	0.95
Oral Communications	284	5.13	1.12
Written Communications	284	5.10	1.30
Arithmetic Skill	282	4.82	1.05
Reading	284	5.37	1.09
Skill in Observing Behavior	284	5.12	1.13
Self-Defense/Physical Restraint of Others	278	4.11	1.38
Social Skills	284	5.07	1.26
Motivation/Dependability	284	5.14	1.35
Stress Tolerance	284	4.95	1.18
Dominance/Self-Esteem	284	4.98	1.26
Intelligence and Judgment	284	4.98	1.22
Personal Organization	280	4.92	1.44
Tolerance of Negative Aspects of the Job	281	4.60	1.18
Flexibility	283	4.83	1.23
Overall Job Performance	278	4.96	1.17

[a]Ratings made on a 7-point scale.

The distribution of scores is somewhat negatively skewed, indicating that more scores are clustered toward the higher ratings than would be expected in a normal distribution.

Inspection of the intercorrelations of the 32 ratings scales indicated that ratings within categories were not more highly interrelated than were ratings across categories. For example, the average intercorrelation of the 10 behavioral summary scales is .47 and the average intercorrelation of the 21 worker characteristic scales is .44. The average correlation of the 10 behavioral summary scales with the 21 worker characteristic scales is .42 indicating that neither of these rating formats contributed significant unique method variance to the ratings data. This finding allows the ratings to be combined across all scales as is discussed in the next section.

Developing an overall criterion. The principal goal of the criterion analyses was to develop a single measure of job performance that would serve as the criterion for subsequent validation analyses. A criterion-related research design requires that there be a single criterion against which the validity of the overall selection battery is judged.

Inspection of the data in Table 6.1 indicated no reason to give precedence to any one of the individual rating scales as a measure of overall effectiveness. By itself, the 7-point overall effectiveness scale was not appropriate as the principal criterion for two reasons. First, individual scales do not provide as reliable measures as do combinations of scales, nor do they provide sufficient variation in criterion scores to permit an accurate assessment of the magnitude of the relationship between predictor and criterion. Second, the overall effectiveness scale may have meant different things to different raters. For example, rater A may have placed more weight on Oral Communications while rater B may have weighted Flexibility more heavily. Thus, incumbents might not have been weighted on the same basis from rater to rater. Because the overall effectiveness scale was inappropriate and because no other scale stood out as deserving special attention, an overall effectiveness score was constructed by averaging across all 32 rating scales.

The overall effectiveness criterion (hereafter called overall criterion) was computed by: a) standardizing each rating scale (T-scores; mean = 50.00, SD = 10.0); b) computing an average rating for each scale type (i.e., behavioral summary scales, worker characteristic scales, and overall rating) and; c) taking the average rating across the three averages. As long as an incumbent received ratings on 90% or more of the 32 scales, missing ratings were treated by averaging across

those scales for which ratings were made. Standardizing the scales prior to computing the overall criterion resulted in each scale contributing equally to the final average. The above method resulted in an overall criterion with a mean of 50.0 and a standard deviation of 7.5 (based upon 284 cases). Table 6.2 presents the correlation of this overall criterion with each of the individual rating scales. The average of the 10 behavioral summary scales correlated .94 with the overall criterion, and the average of the 21 worker characteristic scales correlated .91 with the same criterion. The magnitude of these correlations is due in part to the fact that the behavioral summary and worker characteristic scales each comprise one-third of the overall criterion. Thus, the correlations are inflated by the fact that they, in large part, represent the relationship of an average with itself. The average of the behavioral summary scales correlated .84 with the average of the worker characteristic scales. The average of the behavioral summary scales and the average of the worker characteristic scales correlated .81 and .75, respectively, with the single overall effective scale.

An internal consistency approach was taken to compute the reliability of the overall criterion. Criterion reliability was found to be .96. Originally, multiple ratings were to be collected for each incumbent in the validation sample. However, discussions with department coordinators indicated that it would be very difficult to locate even two supervisors who were sufficiently familiar with an incumbent's performance to provide accurate ratings. The internal consistency approach used is the same that was described in Chapter 3 to estimate the reliability of other ratings data. This approach is somewhat susceptible to overestimating the ratings reliability due to a tendency on the part of raters to form a global impression of incumbents' performance and not to distinguish between different levels of performance on the different ratings scales (halo error). The rating instructions warned raters against this tendency and requested that care be taken in making the ratings. Nevertheless, the reliability reported above will slightly overestimate the true reliability of the ratings.

Ratings reliability is important in validation research only because it places a limit on the validity of the predictors. The less reliable the ratings, the lower the maximum possible validity of the predictors. Reliability estimates are commonly used to correct validity coefficients for criterion unreliability. Such corrections always increase the magnitude of the estimated validity, with the degree of increase being inversely proportional to the magnitude of the reliability estimate. Because the method used to estimate the reliability of the ratings slightly overestimates that reliability, it is actually a conservative estimate. That is, its use to correct the validity estimate for criterion unreliability will not cause a large increase in the magnitude of the validity coefficient.

Table 6.2

Correlation of 32 Job Performance Ratings Scales with the Overall Criterion

Scale	Corr. w/overall criterion[a]
Behavioral Summary Scales	
Investigating	.77
Making Recommendations	.75
Handling Paperwork/Writing Reports	.67
Est. Relationship with New Probationers	.68
Monitoring and Enforcing Compliance	.64
Dealing with Crises/Counseling	.65
Informing/Advising Non-Probationers	.70
Interacting with Attorneys/Judges	.67
Interacting with Other Staff/Agencies	.69
Commitment to the Profession, Dept.	.70
Worker Characteristic Scales	
Department-Specific Knowledge	.71
Knowledge of Specific Laws	.65
Knowledge of Legal System, Terminology	.60
Knowledge of Psychology	.60
Knowledge of Chemical Use, Abuse	.44
Interviewing and Counseling Skill	.73
Safety, First Aid	.30
Oral Communications	.67
Written Communications	.68
Arithmetic Skill	.41
Reading	.57
Skill in Observing Behavior	.71
Self-Defense/Physical Restraint of Others	.39
Social Skills	.64
Motivation/Dependability	.73
Stress Tolerance	.68
Dominance/Self-Esteem	.66
Intelligence and Judgment	.76
Personal Organization	.69
Tolerance of Negative Aspects of the Job	.58
Flexibility	.71
Overall Job Performance	.94

[a]The number of cases varies from approximately 276-284; all r's are significant at alpha = .05.

Search for additional criteria. Apart from the development of an overall criterion, the criterion analyses had as an additional goal the development of criteria relating to more specific aspects of job performance. To the extent that such specific criteria were predictable using one or more of the tests in the experimental selection battery, agencies needing to assess those specific areas of job performance would have valid instruments at their disposal.

The search for additional criteria was conducted via factor analysis of the ratings data. The matrix containing the criterion scale intercorrelations was divided into three submatrices: one containing the intercorrelations of the behavioral summary scales, the second containing the intercorrelations of the worker characteristic scales, and a third containing the intercorrelations of the behavioral summary and the worker characteristic scales. None of the matrices contained the overall effectiveness rating scale. An iterative principal components analysis, followed by varimax rotation to simple structure, was performed for each submatrix as is further described.

Analysis of each submatrix began with the extraction of a single factor. This was followed by additional analyses extracting two, three, . . . N, factors at each step. After each additional factor was extracted, the factor loading matrix was examined to determine whether the resulting solution made sense in light of what was known about the probation officer job. Additionally, the traditional factor solution criteria of eigenvalue magnitude and variance accounted for were monitored. The iterative procedure continued until the extraction of the next factor either made no sense, or drastically exceeded typical rules of thumb for acceptable factor solutions (e.g., eigenvalue less than 1.0).

A three-factor solution was determined to be most appropriate for the behavioral summary scales submatrix and a five-factor solution was accepted for the worker characteristics submatrix. Because the two solutions were judged to have considerable overlap, the factor analysis involving both behavioral summary and worker characteristic scales was taken as the most appropriate basis on which to construct additional, specific criteria. Descriptions of the resulting factors are given in Table 6.3. Figure 6.1 indicates how each of the 31 behavioral summary or worker characteristic scales loaded on the five-factor solution was judged to be most appropriate. It should be noted that the "all or nothing" loadings shown in Figure 6.1 represent the forced simple structure solution that was used to compute scores for each additional criterion.

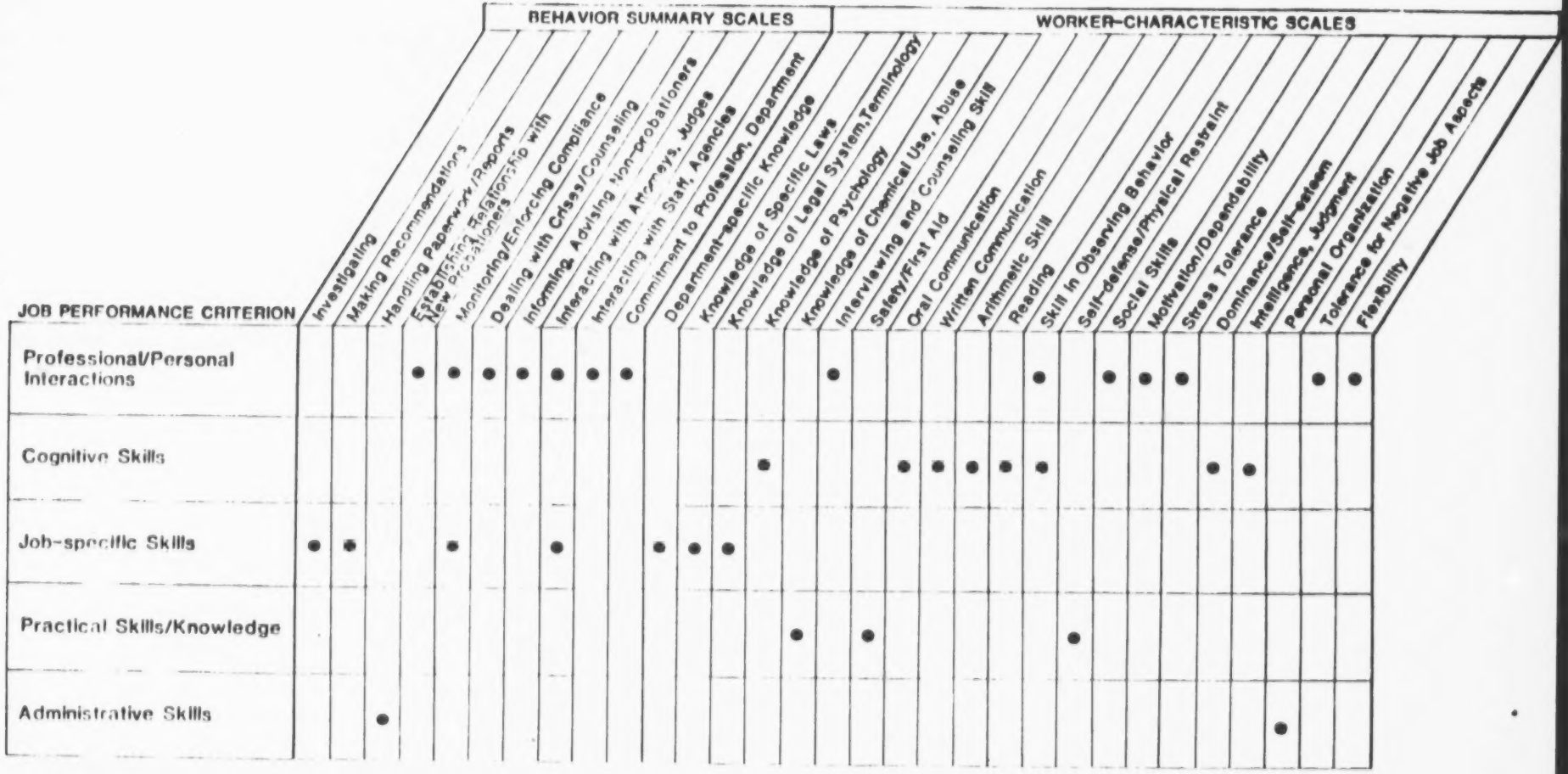
Criterion scores were determined by averaging across the standard scores (T-scores) of the rating scales loading on a particular factor. This method of computing factor scores differs from more commonly used methods in that in the latter, factor loadings are used to weight the contribution of each variable to each factor score. In such an approach every variable contributes to a factor score -- even those variables

Table 6.3

Factor Descriptions of Five Specific Job Performance Criteria

Criterion	Description
Professional and Personal Interactions	This criterion involves skills necessary for harmonious professional and personal relationships with others. These skills involve the ability to interact with attorneys, judges, and other staff and agencies. Interviewing and counseling skills and the ability to deal with crises are important components of this criterion. Also included here are flexibility, tolerance, and motivation.
Cognitive Skills	This criterion revolves around the 3 R's -- reading, writing, and arithmetic. It involves the ability to speak and write well, and requires intelligence, judgment, and general knowledge about psychology.
Job-Specific Skills	This criterion assesses a probation officer's knowledge of the department, specific laws, and the legal system and specific terminology. A successful officer will be able to monitor and enforce compliance and also interact with attorneys and judges.
Practical Skills and Knowledge	This criterion assesses incumbents' knowledge of self-defense techniques, safety and first aid procedures, and chemical use and abuse.
Administrative Skills	This criterion requires a sense of personal organization and the ability to handle paperwork and reports.

Figure 6.1 Simple Structure Factor Loadings of 31 Ratings Scales on Five Specific Job Performance Criteria



● Indicates that the rating scale was used to define the job performance criterion and contributes to an incumbent's score for that criterion.

not considered to be part of a particular factor. The approach used in this project was to assign a rating scale only to the factor on which it loaded most clearly. Factor scores were the unit weighted average of all scales that define a given factor. Due to the nature of factor analysis, these factor scores, although derived from an orthogonal (varimax) rotation, are not uncorrelated.

To summarize, the difference between typical methods of computing factor scores and the method used in this project is that the typical method yields factor scores that are uncorrelated but that are computed using variables not included in the factor definition. The current method yields factor scores based only upon those variables that define each factor, but which are substantially intercorrelated. Table 6.4 presents the descriptive statistics for each of the five additional criteria as well as the intercorrelations of the five factor scores.

Criterion descriptive statistics. The criterion analyses yielded a total of six criteria: a criterion of overall effectiveness and five criteria of the more specific aspects of job performance (Professional and Personal Interaction, Cognitive Skills, Job-Specific Skills and Knowledge, Practical Skills, and Administrative Skills). Table 6.5 presents descriptive statistics for these criteria broken out by incumbent gender and ethnic status (the latter only for those groups represented in sufficient numbers to permit accurate calculations). Female probation officers received higher ratings than male officers on Cognitive Skills and lower ratings than males on all the other criteria. Minority officers were rated lower than nonminority officers on all criteria except Professional and Personal Interactions, where they were rated essentially the same as nonminorities. There was concern that the lower performance ratings of female and minority probation officers on these criteria might have been due to bias by the raters rather than due to true differences in the average performance of these probation officers. This concern led to the calculation of the average criterion scores by the ethnic status of raters as well as the ethnic status of incumbents. Data for gender are given in Table 6.6. Care must be taken not to overinterpret these data as there are very few ratings available from female supervisors. The ratings for male and female incumbents overall were not statistically significantly different from each other. There was, however, a statistically significant difference in the mean ratings assigned by male and female supervisors: Females rated all incumbents lower than did males, regardless of the sex of the incumbents. There was a tendency for female supervisors to rate males lower than they rated females; the opposite was true for male supervisors. However, this interaction effect was not statistically significant and one can only conclude that there is no sex bias in these ratings.

Table 6.4

Descriptive Statistics and Intercorrelations
for Five Factor Score Criteria

Criterion	Avg.	S.D.	Correlation with: 1	2	3	4	5
1. Professional and Personal Interactions	49.9	7.3	--				
2. Cognitive Skills	49.9	7.7	.78	--			
3. Job-Specific Skills and Knowledge	50.1	7.8	.79	.78	--		
4. Practical Skills and Knowledge	49.9	7.7	.48	.49	.51	--	
5. Administrative Skills	49.9	9.4	.60	.56	.61	.30	--

Table 6.5

Descriptive Statistics for the Overall and Specific Job Performance Criteria

Group		Overall	Specific Criteria[a] 1	2	3	4	5
Gender							
Males (175)	Mean	50.2	50.2	49.7	50.3	51.3	50.0
	SD	7.6	7.4	7.9	7.9	7.6	9.5
Females (103)	Mean	49.7	49.5	50.3	49.7	47.5	49.8
	SD	7.3	7.3	7.3	7.8	7.5	9.3
Ethnic Group							
Blacks (32)	Mean	48.6	50.1	47.3	47.9	49.2	48.1
	SD	8.5	7.2	8.8	8.5	6.9	10.5
Hispanics (18)	Mean	49.9	49.7	47.9	48.2	49.9	48.6
	SD	6.4	6.9	6.6	7.4	9.1	9.0
Whites (219)	Mean	50.2	49.9	50.6	50.6	50.1	50.2
	SD	7.5	7.5	7.6	7.8	7.7	9.3

[a] 1 = Professional and Personal Interactions

2 = Cognitive Skills

3 = Job-Specific Skills and Knowledge

4 = Practical Skills and Knowledge

5 = Administrative Skills

Table 6.6

Average Overall Criterion Scores by Supervisor and Incumbent Gender

	Supervisor	
Incumbent	Male	Female
Male		
N	145	19
Mean	50.9	45.8
S.D.	7.1	8.1
Female		
N	87	11
Mean	49.9	49.4
S.D.	7.4	5.6

Ratings data for ethnic groups are given in Table 6.7. Care must be taken not to place too much emphasis on the sparse data in this table. There were no statistically significant differences in the average ratings made by raters from different ethnic groups, nor were there significant differences in the ratings received by incumbents from different ethnic groups. There was no significant statistical interaction between supervisor and incumbent ethnic groups. Taken together, the data in Tables 6.6 and 6.7 give no indication that the mean differences in criterion scores of various subgroups are due to anything besides true differences in the ability of members of those groups to perform the probation officer job tasks.

Predictor Analysis

Predictor analyses were conducted to ensure that all predictor data were accurately collected and scored, to evaluate the 15 tests in terms of their psychometric quality, and to describe the performance of the validation sample on each of the tests.

Data collection and cleaning. Incumbents recorded their test answers on machine readable answer sheets. Upon receipt of the answer sheets, the first step in predictor analysis was to check each sheet to ensure that all marks were dark enough for the optical scanner to "see" and to make sure that the ID number on every answer sheet corresponded to the number printed on the test booklets.

Once checked, the answer sheets were read by an optical mark reader and transferred to the computer where they would be further scored and analyzed. As part of scanning, any answer sheet that contained one or more apparently missing responses was flagged for inspection. This was necessary to ensure that marks too light for the scanner to "see" were not missed. Where necessary, marks were darkened and the answer sheets rescanned.

Scale revisions and final psychometrics. Once all predictor data had been accurately scanned, the 15 tests comprising the experimental selection battery were scored and examined for indications of needed revisions. The goal of scale revision was to achieve a final set of reliable tests for consideration during the validation analyses.

Scale revisions began with the computation of item-total correlations and item difficulty indices. Item-total correlations were computed using the point-biserial correlation and item difficulty was computed as the proportion of incumbents passing each item. Both statistics were computed for every item and distractor. The item statistics were computed using data from all 322 completed experimental selection batteries.

Items for which the keyed alternative correlated negatively with total scores were dropped from the tests. Additionally, typographic errors during the assembly of the experimental selection battery resulted in one item being dropped from each of the Reading, Arithmetic, Vocabulary, and Verbal Reasoning tests.

Item difficulty statistics for the Verbal Reasoning test indicated that few incumbents were successfully completing the last five items on that test and so those items were dropped.

Psychometrics for all tests (after revisions) are given in Table 6.8. Internal consistency (KR-20) reliability coefficients were computed for the cognitive ability tests; published estimates are given for the noncognitive tests. As can be seen, the tests are all sufficiently reliable to be used in pre-employment screening. In addition, the tests were all of moderate difficulty, indicating that they were appropriate for use in the candidate population.

Descriptive statistics. Table 6.9 gives the mean (average score) and standard deviation of each test in the experimental selection battery. Table 6.10 shows the intercorrelations among those 15 tests. Tables 6.11 and 6.12 give the descriptive statistics of each test broken out by incumbents' gender and ethnic status, respectively. Female probation officers performed better on all of the cognitive ability tests than did their male counterparts. On the noncognitive tests, the male officers scored higher than the female officers.

Black and Hispanic probation officers received scores that were considerably below those of white probation officers for all of the cognitive ability tests. Minority officers received higher scores than nonminority officers for all of the noncognitive tests.

Validity and Fairness Analysis

Validity and fairness analyses had as their basic goal an evaluation of the potential of each test to provide valid, fair predictions of probation officer job performance. Valid tests are those for which high scores indicate a higher probability of successful job performance than do low scores. Fair tests are those for which no average score differences occur among gender or ethnic groups except where differences of a similar magnitude occur for the criterion.

Once the validity and fairness of each test was determined, a combination of the tests could be chosen to form the Probation Officer Examination, meeting the project goal of developing and validating entry-level selection procedures for the probation officer position.

Table 6.7

Average Overall Criterion Scores by
Supervisor and Incumbent Ethnic Status

Incumbent	Supervisor		
	Black	Hispanic	White
Black			
N	4	3	23
Mean	52.5	--	51.1
S.D.	9.7		7.6
Hispanic			
N	5	--	12
Mean	54.6	--	49.3
S.D.	3.3	--	5.7
White			
N	7	11	189
Mean	52.0	48.4	50.1
S.D.	9.5	8.7	7.3

Table 6.8

Final Item Statistics for the
Experimental Selection Battery

Test	Number of Items Original	Dropped	Avg. Disc.	Avg. Diff.	rel.[a]
Observational Memory	30	-	.15	.75	.63
Reading	70	1	.27	.82	.88
Arithmetic	35	1	.33	.82	.85
Language Usage	35	-	.26	.81	.80
Vocabulary	35	1	.33	.89	.75
Analogies	35	-	.30	.74	.86
Verbal Reasoning	30	6	.34	.68	.84
Sociability	35	2	.21	.70	.7-.85
Empathy	29	6	.11	.60	.84
Control	24	-	.34	.74	.84
Socialization	49	6	.19	.74	.65-.80
Responsibility	36	1	.18	.69	.65-.85
Stress Reaction	24	-	.48	.76	.90
Dominance	44	5	.26	.67	.65-.80
COIB	58	-	.12	.64	.70

[a]Reliability for noncognitive tests are from published sources; the remainder are KR-20 estimates.

Table 6.9

Descriptive Statistics on 15 Individual Tests for the Research Sample

Test	N[a]	Avg.	Stnd. Dev.
Observational Memory	312	22.53	3.62
Reading	321	56.70	7.98
Arithmetic	321	27.97	4.91
Language Usage	321	28.36	4.53
Vocabulary	321	30.21	3.30
Analogies	321	25.74	5.67
Verbal Reasoning	305	16.37	4.85
Sociability	316	23.23	4.33
Empathy	316	13.84	2.94
Control	314	17.77	4.64
Socialization	315	31.67	3.99
Responsibility	317	24.24	3.82
Stress Reaction	315	18.31	4.67
Dominance	316	26.21	5.45
COIB	308	29.59	3.56

[a]Ns less than 322 are due to tests not being scored for individuals with excessive missing responses.

TABLE 6.10

Intercorrelation of 15 Predictor Tests

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Obs. Memory	-														
2. Reading	.26*	-													
3. Arithmetic	.29*	.50*	-												
4. Language Usage	.23*	.60*	.53*	-											
5. Vocabulary	.10*	.58*	.55*	.62*	-										
6. Analogies	.33*	.59*	.61*	.64*	.56*	-									
7. Reasoning	.25*	.59*	.45*	.45*	.42*	.59*	-								
8. Sociability	.04	.07	.15*	.02	.03	.07	.06	-							
9. Empathy	.02	.11*	.05	.01	.04	.05	.07	.62*	-						
10. Control	-.10	-.04*	-.02	-.01	-.06	-.04	-.07	-.05	-.20	-					
11. Socialization	-.05	-.07	.03	.06	.14*	-.11*	-.06	.14*	.06	.44*	-				
12. Responsibility	-.06	.00	.02	.04	.08	.01	-.06	.27*	.18*	.29*	.38*	-			
13. Stress Reaction	-.04	-.10*	.03	-.12*	-.11*	-.07	-.11*	.33*	.26*	.22*	.33*	.33*	-		
14. Dominance	-.13*	.03	.11*	.06	-.02	.02	-.04	.58*	.51*	-.50	.12*	.42*	.31*	-	
15. COIB	-.01	.00	-.01	-.11*	-.09*	-.09	.02	0.06	-.06	.11*	.40*	.03*	.19	-.08	-

*Statistically reliable correlation at alpha = .05.

Table 6.11

Descriptive Statistics on Individual Tests
by Incumbent Gender

	Males[a]		Females[b]			
	Avg.	S.D.	Avg.	S.D.	t	ES[c]
Observational Memory	22.0	3.3	23.3	3.6	3.13[d]	.38
Reading	56.2	8.0	58.0	7.3	1.94	.23
Arithmetic	28.0	4.8	28.0	4.9	.08	.01
Language Usage	27.5	4.3	29.8	4.4	4.50[d]	.54
Vocabulary	29.9	3.5	30.8	2.8	2.42[d]	.29
Analogies	24.8	5.5	27.4	5.1	4.11[d]	.17
Verbal Reasoning	16.1	4.7	16.9	4.6	1.27	.16
Sociability	22.9	4.4	23.6	4.3	1.29	.15
Empathy	13.7	2.9	13.9	3.0	.56	.07
Control	17.7	4.5	17.8	4.3	.19	.02
Socialization	31.7	4.1	31.7	3.9	-.15	-.02
Responsibility	23.8	3.8	25.0	3.7	2.64[d]	.32
Stress Reaction	18.3	4.8	18.2	4.6	.26	-.03
Dominance	25.9	5.6	26.6	5.4	1.06	.13
COIB	30.2	3.8	28.7	3.0	3.38[d]	-.41

[a] N = 180, approximately

[b] N = 110, approximately

[c] ES = Effect size (standardized mean difference); negative values

[d] statistically significant at alpha = .05

Table 6.12

Description Statistics on 15 Individual Tests
by Incumbent Ethnic Status

Test	Blacks[a] Avg.	S.D.	Hispanics[b] Avg.	S.D.	Whites[c] Avg.	S.D.	F	ES[d]
Obs. Memory	21.3	3.7	21.2	3.3	22.9	3.4	3.25[e]	-.49
Reading	54.2	9.6	51.0	9.8	57.8	7.1	5.38[e]	-.63
Arithmetic	25.2	4.6	24.0	5.7	28.8	4.5	8.60[e]	-.87
Language Usage	26.8	4.1	25.B	5.7	28.9	4.3	4.43[e]	-.57
Vocabulary	29.3	3.2	27.1	5.0	30.7	2.9	7.52[e]	-.69
Analogies	22.2	6.0	22.0	5.1	26.7	5.2	9.04[e]	-.87
Verbal Reasoning	13.2	5.2	11.4	5.3	17.3	4.1	13.52[e]	-1.12
Sociability	24.4	3.9	23.1	3.9	23.0	4.5	.84	.20
Empathy	14.1	3.3	13.6	3.3	13.8	2.9	.86	.02
Control	18.7	3.7	17.9	4.0	17.6	4.5	4.05[e]	.18
Socialization	34.2	3.4	31.6	3.4	31.4	4.0	6.83[e]	.48
Responsibility	25.5	3.3	23.4	3.8	24.2	3.9	1.73	.16
Stress Reaction	19.2	4.5	20.1	3.4	18.0	4.7	1.83	.33
Dominance	27.5	6.2	26.4	4.3	26.0	5.5	.95	.22
COIB	30.1	3.7	29.2	3.7	29.6	3.6	.35	.05

[a] N = 30

[b] N = 20

[c] N = 232

[d] Standardized mean difference contrasting blacks and Hispanics to whites; negative value indicates lower average scores by blacks and Hispanics.

[e] Statistically significant at alpha = .05

An additional goal of the validity and fairness analyses was to determine whether any of the five specific job performance criteria were differentially predictable in a fair and valid manner, either by individual tests or by a different combination of tests than that chosen to predict overall job performance.

Adverse impact and test fairness. Tests for which two or more groups receive scores that are statistically different on average are said to discriminate between the groups. Used in this limited sense, the term discrimination simply indicates that there is a statistically reliable difference between the groups (i.e., a difference likely to be found every time the groups are tested). The presence of group differences for a selection procedure does not automatically disqualify that procedure however. The Uniform Guidelines on Employee Selection (1978) discuss the relationship between two important concepts related to the fair use of selection procedures. The first of these is adverse impact and the second is test fairness.

Adverse impact (as defined by the 4/5ths rule of thumb) exists whenever the selection ratio for members of a protected group is 4/5ths (80%) or less than that of the group with the largest selection ratio. Unfairness is defined in the guidelines as occurring when members of one group obtain lower scores on a selection procedure and when similar differences are not found for measures of job performance. The Guidelines also note that test fairness must be investigated (if technically feasible) whenever a selection procedure results in adverse impact, but not if adverse impact is absent.

From the perspective of selection procedures, discrimination in the limited sense described above can be either fair and therefore permissible, or unfair and therefore not permissible. In evaluating a selection procedure the first step, then, is to determine whether adverse impact exists. Adverse impact is a joint function of the magnitude of group differences and the overall selection ratio used during selection. Even very large group differences may not result in adverse impact if the selection ratio is high enough.

Many different models have been offered for use in distinguishing fair from unfair discrimination. One model (Cleary, 1968) holds that group differences in predictor scores can only be considered fair if differences of the same magnitude occur for the criterion measure. In statistical terms, this translates into a test of the similarity of regression equations computed separately for each subgroup involved in the fairness analysis. Selection decisions for any valid test or procedure depend upon the use of a single regression equation to predict the expected performance of all candidates. A selection procedure is unfair to the members of any group whose performance is systematically underpredicted by a common regression equation.

The standardized mean difference between groups (effect size) is an important statistic in this determination of test fairness. In the Cleary model, tests biased toward minority groups (large negative effect size) will only be fair if a difference of comparable magnitude exists for criterion scores. From the perspective of adverse impact, selection procedures with large negative effect sizes require that very low cutoff scores be set in order that sufficient numbers of protected group members be selected. Effect sizes indicating bias against protected groups warn of the necessity of further analyses to determine whether adverse impact is present, and if adverse impact exists, whether the test is fair to protected groups.

Validity and fairness of the individual tests. Following all necessary revisions, each of the tests in the experimental selection battery was correlated with the overall criterion. At the same time, average scores and associated descriptive statistics were computed for all members of the final validation sample as well as separately by incumbent gender and ethnic group. These descriptive statistics were reported in Tables 6.11 and 6.12. Table 6.13 presents the validity coefficients for the 15 tests in the experimental selection battery. As can be seen in this table, for three tests (Arithmetic, Analogies, and Stress Reaction) the validity estimates were statistically significant at the .05 level. Of these three, Arithmetic stands out as the best single predictor of job performance when magnitude of the validity coefficient is the sole criterion of "best predictor". The only other tests that demonstrated any promise of providing accurate indications of probation officer job performance were Reading and Control.

Tables 6.11 and 6.12 presented descriptive statistics and fairness statistics by gender and ethnic status, respectively, for each test. For each test, means and standard deviations by subgroup are given, along with the appropriate statistic for determining whether differences between groups are statistically significant. Finally, an index of the magnitude of the difference between groups is reported. This index, effect size, gauges the distance between group means relative to the pooled variation in scores. Conceptually, effect size estimates are interpreted in the same way as are z scores. For example, an effect size of .30 indicates that the mean of one group lies .30 pooled standard deviations above the mean of the second group and, therefore, that 62% of the members of the first group receive scores greater than the average score of the second group. In groups with equal means, 50% of persons in either group receive scores above the mean of the other group.

Table 6.13

Correlations of 15 Individual Tests
with Overall Job Performance

Test	Corr. w/job perf.	p[a]
Observational Memory	.04	.26
Reading	.09	.07
Arithmetic	.16	.00
Language Usage	.04	.27
Vocabulary	.05	.21
Analogies	.12	.02
Verbal Reasoning	.05	.21
Sociability	-.04	.23
Empathy	-.01	.43
Control	.09	.07
Socialization	-.02	.35
Responsibility	.07	.14
Stress Reaction	.11	.03
Dominance	-.01	.43
COIB	.01	.46

[a]Probability that a correlation this large or larger would be obtained due to chance. Probabilities below .05 are considered statistically significant.

The data in Table 6.11 show that none of the tests that were significantly correlated with the overall criterion (or nearly so) had statistically significant mean gender differences. Thus, those tests shown to be valid predictors of job performance are also fair predictors for males and females.

Fairness data for the three ethnic groups with large enough samples to permit calculation of reliable statistics are given in Table 6.12. These statistics are quite similar to those reported for gender except that a one-way analysis of variance rather than a t test was used to determine whether the differences among groups were statistically significant. The effect size measures for ethnic groups contrast the combined group of black and Hispanic officers to white officers.

Two of the tests for which statistically significant validity estimates were found (i.e., Analogies and Arithmetic) also displayed statistically significant mean differences across the three ethnic groups. In both cases, these differences favored whites. There were no significant group differences in Stress Reaction test scores. Reading and Control also showed some promise as valid predictors of performance. Mean scores on Control significantly favored blacks and Hispanics; scores on the Reading test, however, were significantly higher for whites. Comparison with the data in Table 6.5 shows that the mean differences among ethnic groups on these three cognitive tests are much larger than the corresponding mean differences on the overall criterion. Thus, most of the tests showing promise as valid measures of job performance were not fair predictors for the three ethnic groups.

Developing a final selection battery. Following the evaluation of the fairness and validity of each of the 15 individual tests, it was necessary to determine which combination of those tests would be recommended for use as the Probation Officer Examination. Several constraints influenced the choice of tests to be included in the final battery. Principal among these was the necessity that the battery be valid and either incur no adverse impact or provide fair predictors of the performance of candidates from all groups. Of considerable additional importance was that the final selection procedures assess candidates' potential in all important job areas, that the final set of tests require no more than 2 to 2 1/2 hours to administer, and that scoring be as simple as possible so as not to compromise the operational validity of the final battery.

A core set of tests was selected from those that made up the experimental selection battery according to the following procedure. First, those tests (Arithmetic, Analogies, and Stress Reaction) that were significantly correlated with the overall criterion were included in the final selection battery in their entirety. The entire Control test was also added because of its correlation with the criterion and because of its positive effect size (i.e., mean score difference favoring minorities). Thus, the core set of tests included the original Arithmetic, Analogies, Control, and Stress Reaction tests. A total score

was then computed for each probation officer incumbent by summing his/her scores on the four selected tests. This combination of tests was found to be a valid predictor (r with overall criterion = .23, $p < .05$). This core resulted in a sizeable mean difference favoring whites (effect size = -.57), suggesting that adverse impact would be present.

Several procedures are available that permit the use of apparently unfair tests in a manner that does not result in adverse impact against protected groups. For example, separate selection lists can be constructed and candidates can be selected from both lists in numbers that do not violate the 4/5ths rule of thumb. Another method is to adjust test scores by adding a constant value to some scores, thus removing the mean differences among groups. As neither of these methods were administratively feasible for the Board of Corrections, the following method was used to expand the final selection battery while maintaining or increasing its validity and decreasing the magnitude of the mean differences among ethnic groups.

Because none of the remaining tests from the experimental selection battery could be used in its entirety without increasing the bias toward minority groups, a search was undertaken at the item level. Because the cognitive ability tests were highly intercorrelated, and because they were the principal source of mean differences favoring whites, the pool of items to be searched was narrowed to just those items from the noncognitive tests.

The final validation sample was randomly split into halves and the effect size of each item (blacks and Hispanics versus whites) was computed for every item in both samples. All items that had positive item validities and that favored minorities in <u>both</u> samples were selected for use in the final battery.

This procedure identified 31 items drawn from a maximum possible 231 items from the noncognitive tests. Table 6.14 indicates the source of the items selected using the methods described above. Predictor scores were recomputed for all incumbents by summing across the items that comprise the final selection battery. These scores were found to correlate .23 with the overall criterion, indicating that the battery was a valid measure of job performance.

Fairness data for the Probation Officer Examination are given in Table 6.15. Females score higher on the examination than do males; this difference is statistically significant. The ethnic group differences are not statistically significant. The effect size for gender on the predictor is .24, while the corresponding effect size on the overall criterion is -.07. For ethnic groups, the predictor effect size is -.19, while the criterion effect size is -.14. That the mean ethnic group differences are not statistically significant indicates that they are no larger than would be expected due to chance differences that occur between any groups of similar size. Nevertheless, the predictor effect size for ethnic groups is larger than the corresponding criterion effect size. This indicates the need to examine whether the Probation Officer Examination results in adverse impact against minority candidates.

As discussed previously, effect size is one of two critical factors in determining whether a test will violate the 4/5ths rule of thumb, the other factor being the cutoff score chosen. Score distributions for minority and nonminority incumbents were checked for normality. Scores for both groups were determined to be normal. In addition the variances of the two score distributions were found to be statistically equivalent. These analyses demonstrated that minority and nonminority score distributions differed only in tests of their means. This finding permitted the computation of the expected adverse impact that would result from various overall selection ratios.

For the Probation Officer Examination, as long as the overall selection ratio was greater than or equal to .22 (raw score of 129), no violation of the 4/5ths rule occurs. A selection ratio of .22 is more stringent (only 22% of all applicants pass) than is recommended for an examination of this type. Even if a cutoff score of 129 was set, use of the Probation Officer Examination to "select" candidates from the final validation sample would not have resulted in adverse impact against minority group members. Assuming that the effect size for ethnic groups remains near -.19 in future applicant groups, and further assuming that minority and nonminority groups have approximately equal score distributions, in order to avoid adverse impact, selection ratios must be kept above .22, a value well above that recommended for use with this battery (See Chapter 7). As no adverse impact was found for minority probation officers, no further analyses were necessary.

Adverse impact analyses for females revealed that there were no overall selection ratios that would lead to adverse impact against women. Because no adverse impact against female probation officers was found, no further analyses were necessary.

<u>Differential predictability of specific criteria</u>. Correlations of the 15 tests and the final predictor with five specific criteria of job performance are shown in Table 6.16. As can be seen, there were several individual tests that provided valid predictions of one or more of the specific criteria. As can also be seen, the Probation Officer Examination was nearly as valid or more valid a predictor as were any of the individual tests. Tables 6.11 and 6.12 reported the likely adverse impact of the individual tests against female and minority candidates. Because the individual tests invariably result in considerable adverse impact, none of them would be both fair and more valid than would the Probation Officer Examination. It was therefore concluded, that the specific criteria were most predictable using Probation Officer Examination scores, so no additional predictors were investigated.

Table 6.14

Source of Items Selected to
Augment the Final Selection Battery

Source	Items Selected[a]
Sociability	7
Empathy	4
Responsibility	11
Socialization	9
Dominance	7
COIB	2

[a]Number selected sums to 38 due to seven items which are scored on two scales.

Table 6.15

Gender and Ethnic Group Statistics for Probation Officer Examination Scores

Group	N	Mean	S.D.	F	ES
Males	179	117.4	12.2		
Females	110	120.3	11.9	3.91[a]	.24
Blacks	33	117.2	13.0		
Hispanics	18	116.1	12.7		
Whites	228	119.1	11.9	1.54	-.19

[a]Statistically significant at alpha = .05.

Table 6.16

Correlations of Individual and Final
Predictors with the Five Factor Criterion Measures

Predictor	Factor Criterion[b] 1	2	3	4	5
Observational Memory	-00	09	05	04	01
Reading	06	07	16[a]	-02	14[a]
Arithmetic	13[a]	18[a]	19[a]	06	14
Language Usage	-00	11[a]	09	-05	14[a]
Vocabulary	03	11[a]	13[a]	-00	14[a]
Analogies	08	21[a]	19[a]	02	17[a]
Verbal Reasoning	-00	10	12[a]	-02	07
Sociability	-02	01	-01	02	-13[a]
Empathy	05	06	-01	05	-13[a]
Control	05	01	07	-03	22[a]
Socialization	-02	-09	-03	-07	03
Responsibility	09	12[a]	05	-04	-09
Stress Reaction	12[a]	06	09	02	00
Dominance	-02	08	05	08	-13[a]
COIB	-01	-06	03	10	02
Final Predictor	18[a]	26[a]	24[a]	07[a]	17[a]

Note. Decimal points omitted.

[a]Statistically significant at alpha = .05

[b]1 = Professional and Personal Interactions
2 = Cognitive Skills
3 = Job-Specific Skills and Knowledge
4 = Practical Knowledge
5 = Administrative Skills

CHAPTER SEVEN: DISCUSSION AND RECOMMENDATIONS

The Probation Officer Examination is a fair and valid indicator of potential to successfully perform the probation officer job. PDI therefore recommends that the Board of Corrections adopt this selection instrument (or some other equally valid and fair measure of the same abilities and personal characteristics) as part of the entry-level selection procedures for agencies participating in the STC program. The Probation Officer Examination is not intended to serve as the sole or even as the primary basis for selection decisions. Its judicious use, in combination with other sources of information about candidates' potential (e.g., application appraisals, interviews, educational achievement, etc.), and according to the constraints set forth below will, however, increase the probability that selected candidates will prove to be successful probation officers.

The remainder of this chapter provides a description of the format and content of the Probation Officer Examination and sets forth the more important constraints governing the use of the examination as one source of information about candidates' potential. Recommendations are made concerning: a) scoring the examination; b) the data collection and documentation necessary to supplement the norms collected in this project; and c) and the efforts necessary to document the continued validity and fairness of the examination.

Exam Format and Content

The Probation Officer Examination consists of a single test booklet containing items assessing candidates' cognitive and noncognitive knowledges, skills, abilities, and personal characteristics. The examination contains 150 items and has a total testing time, including time for instructions, of approximately 1 1/2 hours.

Two cognitive abilities are measured in Parts 1 and 2. Part 1 (Analogies) is a measure of how well candidates understand the relationships between words, objects, and concepts. The test consists of 35 analogies problems. Each is in the form of a sentence with the first and last words left out. Five pairs of words are given for each problem. The candidate is asked to choose the pair of words that completes the sentence and makes the analogy true.

Part 2 contains 35 items that measure candidates' facility with the arithmetic operations of addition, subtraction, multiplication, and division. Candidates are to apply these operations to simple multiple choice problems involving whole numbers, fractions, decimals, and percentages.

Part 3 the Probation Officer Examination contains 80 items that assess several personal characteristics found (in the job analysis) to be important components of successful probation officer performance. These items were drawn from scales of the California Psychological Inventory and the Multidimensional Personality Questionnaire. Following are the personal characteristics measured by the items. These descriptions borrow heavily from descriptions of the scales where the items originally appeared.

The MPQ Control and Stress Reaction scales were used in their entirety in the Probation Officer Examination; each of these scales contains 24 items. Another 20% of the items in Part 3 were drawn from the CPI Socialization and Responsibility scales. These CPI scales, along with MPQ Control, reflect a tendency to take a cautious, planful approach to life. Persons who receive high scores on these scales are described by others as being rule-respecting, self-controlled, responsible, and preferring to plan their activities rather than play things by ear. Another personal characteristic measured is dominance (CPI Do). Dominant persons are those who behave in an ascendant manner, who are self-confident, who can take the initiative and exercise leadership in a self-assured manner. High scorers on the MPQ Stress Reaction Scale are those people who are able to put fears and worries out of their minds easily, and who are not troubled by emotional turmoil or feelings of guilt.

The remaining items in Part 3 were drawn from the CPI Sociability and Empathy scales. In general, high-scoring candidates are those who are outgoing and participative, and who are patient and affiliative by nature.

Care should be taken not to overinterpret Probation Officer Examination scores. Individual items were selected for inclusion on the basis of their relationship to the overall criterion (i.e., their criterion-related validity) and their effect on the mean group differences in scores. They were not selected to yield meaningful interpretations regarding underlying psychological constructs. Thus, utility of these items lies in how well they predict, as a set, the overall criterion scores of probation officer incumbents.

Table 7.1 contains descriptions of the noncognitive scales from which these items were selected.

The Probation Officer Examination is a power rather than speed measure. The time limits suggested for the examination are such that most candidates will have ample opportunity to complete all items. The examination equally weights the cognitive and noncognitive items and is scored by counting the number of items that a candidate answers correctly.

Table 7.1

Description of the Noncognitive Tests
Contributing Items to the Probation Officer Examination

Test	Description
CPI Dominance	A subscale of the California Psychological Inventory that identifies individuals who behave in a dominant, ascendant manner and who, in interpersonal situations, can take the initiative and exercise leadership in a forceful, self-confident manner.
CPI Sociability	A subscale of the California Psychological Inventory that indicates a person's tendencies to be outgoing and participative and to seek out and enjoy social contact without being shallow or superficial.
CPI Responsibility	A subscale of the California Psychological Inventory measuring a person's tendency to believe that life is best governed by reason and to behave in a responsible, self-disciplined manner.
CPI Socialization	A subscale of the California Psychological Inventory measuring a person's tendency to be socialized, rule-respecting, prudent, and responsive to what others feel and think.
MPQ Stress Reaction	A subscale of the Multidimensional Personality Questionnaire that indicates persons who are able to put fears and worries out of their mind easily; quickly get over upsetting experiences; and who are not troubled by emotional turmoil or feelings of guilt.
MPQ Control	A subscale of the Multidimensional Personality Questionnaire indicating a person's tendency to be reflective, cautious, careful, rational, and sensible, and who likes to anticipate and plan activities and events.
Empathy	A series of items drawn from the California Psychological Inventory that identifies persons who are described as being "characterized by a patient and forebearing nature, by affiliative and socially ascendant tendencies and by liberal and humanistic political and religious attitudes."

Using the Probation Officer Examination

The Probation Officer Examination is intended to be used only as described in the User's Manual. This manual provides general information concerning the physical settings recommended for testing as well as verbatim proctor's instructions to be followed every time the examination is administered. A Candidate Information Booklet has been prepared that describes the nature and purpose of the examination. No candidate should be tested without having received the Candidate Information Booklet in time to review the information it contains.

The Probation Officer Examination is a fair and valid indicator of candidates' potential to perform the probation officer job but only if it is administered and scored according to the guidelines in the User's Manual. The validity and fairness of any other use of the examination cannot be guaranteed. Persons wishing to use the examination in a manner other than that described in the User's Manual must document for themselves the validity of the examination for that use. All items in the Probation Officer Examination are copyrighted and permission to use the examination in part or in whole must be obtained from the Board of Corrections, Consulting Psychologists Press and the University of Minnesota Press.

Exam Security

Security of the test keys is of the utmost importance. If the security of the keys is breached, the validity of the entire examination is compromised. The governing principle of key security is: The fewer people to have access to the keys the better. The best possible arrangement would be to have all tests scored by the Board of Corrections and to allow only one or two persons access to the test keys. If central test scoring is not possible, the next best arrangement would be to allow only one person from each agency access to test keys. This person would be responsible to the Board of Corrections for the maintenance of test security.

Test keys should not be reproduced indiscriminately. All copies of the keys should be numbered and an accurate record should be kept of their location.

Security of test keys also depends upon the security of the examination booklets. Test booklets should be numbered and accurate records should be kept by the Board of Corrections as to which agencies have which booklets.

Periodic (semi-annual) inventories should be made to ensure that all testing materials can be accounted for and remain secure. If examination materials are lost or destroyed, the Board of Corrections must be informed immediately so that the appropriate action can be taken.

Scoring and Data Collection

Detailed instructions for scoring the Probation Officer Examination are contained in a separate confidential report. Briefly, scoring is accomplished via a set of scoring templates. Candidates will record their responses on an answer sheet. To score the examination, the appropriate scoring template is placed over the answer sheet and the number of responses visible through the template are counted. Parts 1-3 will be scored separately and overall scores will be formed by summing the Part 1 - Part 3 subscores. Space will be provided on the answer sheets to record candidates' scores. Examination answer sheets are designed to permit easy transfer to an electronic scoring systems.

Table 7.2 contains the frequency distribution of incumbents' examination scores. These data show the number and proportion of candidates who received various final scores. Also shown are the actual and expected percentile equivalents of the scores. The actual percentile equivalents are simply the proportion of incumbents who received a score at or below a given value. Expected percentile equivalents reflect the proportion of incumbents who would score at or below a given score based upon a normal distribution of test scores with the same mean and standard deviation as were obtained in the incumbent sample. Because these data are based on the scores of incumbents, they may slightly overestimate the proportions of applicants expected to obtain any given score.

The data in Table 7.2 are the norms that PDI recommends be used to guide test scoring and interpretation (as discussed below) for initial administrations of the Probation Officer Examination. It is important that the Board of Corrections record (at a minimum) the race, gender, and final score of every applicant tested with the Probation Officer Examination. These data are essential for evaluating the appropriateness of the recommended minimum acceptable score and for monitoring the examination's validity and fairness. PDI recommends that the Board of Corrections also retain copies of all applicants' answer sheets so that item level analyses of the Probation Officer Examination can be undertaken as necessary. If examinations are scored by individual agencies (not recommended), copies of all testing materials should be sent to the Board of Corrections. Agencies should retain copies of all testing materials for a minimum of one year, as these materials are likely to be subpoenaed should the examination be challenged in court.

The Board of Corrections is strongly encouraged to conduct annual audits of the Probation Officer Examination to, at a minimum, check scoring accuracy, evaluate the current recommended minimum acceptable score, and evaluate gender and ethnic differences in test scores. It is also recommended that the examination be revalidated against a job performance criterion after it has been operational for two years and every three to five years thereafter unless otherwise indicated by the annual audits.

Table 7.2

Probation Officer Examination Score Distribution

Raw Score	N	Percentile Equivalent	
		Actual	Expected
148	314	99	99
139	308	98	95
135	291	93	90
132	277	88	85
129	252	80	80
127	236	75	75
125	209	67	70
123	192	61	65
122	184	59	60
120	159	51	55
118	142	45	50
117	133	42	45
115	116	37	40
114	108	34	35
112	90	29	30
110	80	26	25
108	68	22	20
105	51	16	15
102	38	12	10
98	26	8	5
89	10	3	1

Score Interpretation and Cutoff Scores

Methods used to set cutoff scores and a recommended minimum acceptable score are contained in a separate report available from the Board of Corrections. Briefly, this report determines the probability that candidates scoring within several score ranges will be adequate or better performers. The recommended minimum acceptable score is that for which there is a 50-50 chance of being an adequate or better performer.

Although a minimum acceptable score has been recommended, no statewide cutoff was set. Choosing an appropriate cutoff score requires consideration of several factors that will vary from agency to agency. This makes the establishment of a statewide cutoff impossible. The cutoff score report discusses the factors that influence the setting of cutoff scores and provides guidelines for agencies to follow in establishing their own cutoff scores. This will allow individual agencies considerable flexibility in determining how best to use the Probation Officer Examination.

References

American Psychological Association. (1974). Standards for Educational and Psychological Tests. Washington, DC: Author.

American Psychological Association, Division of Industrial-Organizational Psychology. (1980). Principles for the Validation and Use of Personnel Selection Procedures. (2nd ed.). Berkeley, CA: Author.

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Equal Employment Opportunity Commission. (1978). Uniform Guidelines on Employee Selection Procedures. Federal Register, 43, 38290-38315.

Gough, H.B. (1982). Manual for the Correctional Officers' Interest Blank. Palo Alto, CA: Consulting Psychologists Press.

Overall, J.E., & Klett, C.J. (1971). Applied Multivariate Analysis (Ch. 7). New York: McGraw-Hill.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, 86, 420-428.

STANDARDS AND TRAINING
FOR CORRECTIONS PROGRAM

STC

BOARD OF CORRECTIONS

TECHNICAL ADEQUACY REPORT

PROBATION OFFICER

Copyright 1987 California Board of Corrections

CALIFORNIA BOARD OF CORRECTIONS

Standards and Training for Corrections Program

Report on the Technical Adequacy of the Final Validation Sample for Probation Officers

In preparing the Probation Officer Examination for the California Board of Corrections, Personnel Decisions, Inc. has taken into account the following documents concerning selection testing:

- The American Psychological Association, Standards for Educational and Psychological Tests, (1974).
- The APA Division of Industrial-Organizational Psychology, Principles for the Validation and Use of Personnel Selection Procedures, 2nd edition, (1980).
- The Equal Employment Opportunity Commission, Uniform Guidelines on Employee Selection Procedures, (1978).

Report On the Technical Adequacy
Of the Final Validation Sample For
Probation Officers

The original project plan for the PDI-STC Probation Officer Validation Project called for testing a minimum of 500 probation officer incumbents. For reasons to be elaborated herein, the actual number of probation officers in the final validation sample ranged from 284 to 322 (depending upon the analysis being conducted). This report documents the data collection efforts resulting in the final validation sample in order to show: (1) that every effort was made to collect usable data from as many incumbents as could reasonably be tested; and (2) that the number of incumbents actually tested is technically adequate to permit the reliable computation of all necessary statistics and analyses.

The data collected in this criterion-related validation project consisted of scores on 15 measures of cognitive and noncognitive skills and abilities (predictors), and job performance ratings in 31 specific areas, plus a rating of overall effectiveness (criteria). For most analyses (e.g., validity, fairness) it was necessary that data for both predictor and criteria be available while other analyses (e.g., norms) required only that predictor data be available.

Predictor Data Collection

Data collection began with telephone canvassing of all agencies that had participated in the job analysis phase of the project. Agencies were contacted to determine their willingness to participate in the testing phase of the project and to get an approximate census of the number of eligible incumbents (i.e., those on the job for six or more months) from each agency. Target samples were set based upon relative agency size, but smaller agencies were generally requested to test a proportionally larger number of their incumbents than were larger agencies. The initial targeted sample called for testing 477 probation officers from 21 probation departments.

Once targeted samples were set, participating departments were contacted by phone to reconfirm their participation in the testing phase of the project and to secure permission to test the targeted number of incumbents. During the follow-up phone calls two agencies informed us that, due to delays in the scheduled testing date, they would be unable to participate in testing. The withdrawal of these agencies reduced the overall targeted sample from 477 to 436.

The remaining 19 agencies were contacted again to see if additional incumbents were available for testing. The agencies committed an additional 12 incumbents to the testing phase, thus raising the total targeted sample to 448.

Because the targeted sample was still well short of the 500 incumbents called for in the project plan, agencies that had participated in the job analysis phase of the project but were not participating during testing were contacted to see if they might be able to supply additional incumbents. No agencies were able to provide additional incumbents for testing.

To further raise the number of incumbents tested, STC consultants contacted a number of agencies that had not previously participated in any phase of the project. Due to the efforts of STC, one agency joined the validation phase of the project and supplied an additional 11 incumbents, raising the total targeted sample to 459.

When it appeared that no additional incumbents were available for testing, materials were distributed to all agencies which had not received them and testing commenced. Unfortunately, not all agencies were able to meet their targeted sample and only 46% of the total targeted sample (296 of 459) was actually tested. The main loss of incumbents occurred in one agency where the participation of incumbents in testing was voluntary. Despite agency solicitation of 800 volunteers, only 19% of the target number of incumbents were actually tested.

A final effort was made to obtain additional incumbents from all eligible agencies throughout California. This effort netted an additional 11 incumbents, raising the total number tested to 322.

Criterion Data Collection

The initial sampling plan called for the collection of job performance ratings for all probation officers tested. Ratings materials were supplied to participating agencies in sufficient quantity to permit rating of all tested incumbents. Coordinators were instructed to select incumbents for testing in a manner designed to minimize the administrative burden of collecting the job performance ratings. Nevertheless, and despite repeated contacts with agencies, ratings data were collected for only 296 of the 322 probation officers for whom predictor data were also available.

Due to test booklets being returned without incumbents' names, incumbents with fewer than six months experience being tested, and ratings made by supervisors who indicated that they did not know the rated incumbents "well at all," the final number of probation officers for whom both predictor and criterion data were available was 284.

Technical Adequacy of the Sample

There are three analyses in particular for which the technical adequacy of the final validation sample must be shown. These are the validation analysis, fairness analysis, and the development of test norms. Each analysis involves one or more specific statistical

tests. Every statistical test involves the testing of a null hypothesis (i.e, no validity, or no difference between groups) versus some, usually unspecified, alternative hypothesis. If an estimate of the magnitude of the effect being tested can be specified (e.g., validity greater than .30, mean difference of .80 standard deviations), the power of the test to detect that effect can be determined. Statistical power is simply the probability (expressed as the number of chances in 100) that the test will reject the null hypothesis if the effect being tested is as large or larger than its estimated value.

The power of all statistical tests is a direct function of the amount of data used to compute the test statistic -- the more data, the more powerful the test. A given sample is technically adequate for a given statistical test if it is of sufficient size to permit a reasonably powerful test of the effect. To determine the power of a statistical test some estimate of the expected magnitude of the effect must be made. The smaller the expected effect, the larger the sample needed to achieve a reasonably powerful statistical test. Statistical power is directly analogous to the power of a microscope where more power is needed to see very small objects than is needed to see relatively larger objects. The technical adequacy of the final validation sample is estimated in the following for the three main analyses of the project.

Validity analyses. The technical adequacy of the final validation sample for validity analyses depends upon the power of the statistical test of the significance of the correlation between predictor and criterion. The correlation coefficient is based upon data from those probation officers for whom both predictor and criterion data are available. In order to determine the power of the statistical test for a sample of 284 some estimate of the expected validity is required. Boehm (1982) reports that the average validity coefficient from 176 studies spanning 20 years was .22. This is the estimate used to test the power of the validity analyses for this project.

Schmidt, Hunter, and Urry (1976) provide a formula to determine the sample size necessary to achieve a desired power for a specified validity coefficient. The problem here is only somewhat different than that posed by Schmidt et al. We want to know the power for a sample of 284 and an estimated validity of .22. The Schmidt et al formula can be converted for use in answering this question. Thus it can be shown (see Appendix) that if the true validity of the final predictor in this project is at least .22, then 96 times out of 100 it will be detected as being statistically significant using a sample of 284. On this basis it is clear that the final validation sample in this project is technically adequate for the purpose of validation analyses.

Fairness Analyses. The technical adequacy of the statistical tests involved in fairness analyses depends upon their ability to detect subgroup differences in average predictor and average criterion scores. As was true for validity analyses, some estimate of the magnitude of the expected effect is necessary to determine the power of the statistical tests. While a straightforward estimate was available for validity, no such easily obtainable estimate exists for gender or race differences.

The final predictor in this study consists of a combination of cognitive and noncognitive measures that has not been previously examined. For general measures of intelligence it is not uncommon to see differences of approximately one standard deviation between minority and non-minority groups, while males and females typically do not differ. For the cognitive component of the examination we might expect to find rather large ethnic group differences favoring whites but no gender differences. The magnitude of typical group differences for the tests that make up the noncognitive portion of the examination is not known. In any case, no sufficiently precise estimate of the expected magnitude of group differences on the Probation Officer Examination is possible.

Without a reasonable estimate of the expected magnitude of group differences no precise determination of statistical power can be made. Figures 1 and 2 show the power of the t test (a statistical test for group differences) plotted as a function of the standardized mean difference between groups. Figure 1 reports the power of the test to detect male-female differences and Figure 2 shows power for detecting ethnic group differences. Note: Although not shown as such, the power curves in Figures 1 and 2 are symmetric about the Y-axis.

The power curves shown in Figures 1 and 2 indicate that for the sample sizes in this project, the statistical test for detecting group differences has sufficient power to detect large group differences and reasonable power to detect moderate group differences (magnitude = .25 - .45 standard deviations), but that the power for detecting small group differences (0.1 - .25 standard deviations) is low.

Fairness analyses depend less upon detecting statistically significant differences between groups than they do upon estimating the magnitude of group differences on predictor and criterion measures. For this reason less powerful statistical tests can be tolerated for fairness analyses (as opposed to validity analyses where statistical significance is the primary concern), as lorg as the sample is of sufficient size to provide a stable estimate of group means and standard deviations.

Figure 1. Statistical power of a t test with 287 df as a function of effect magnitude.

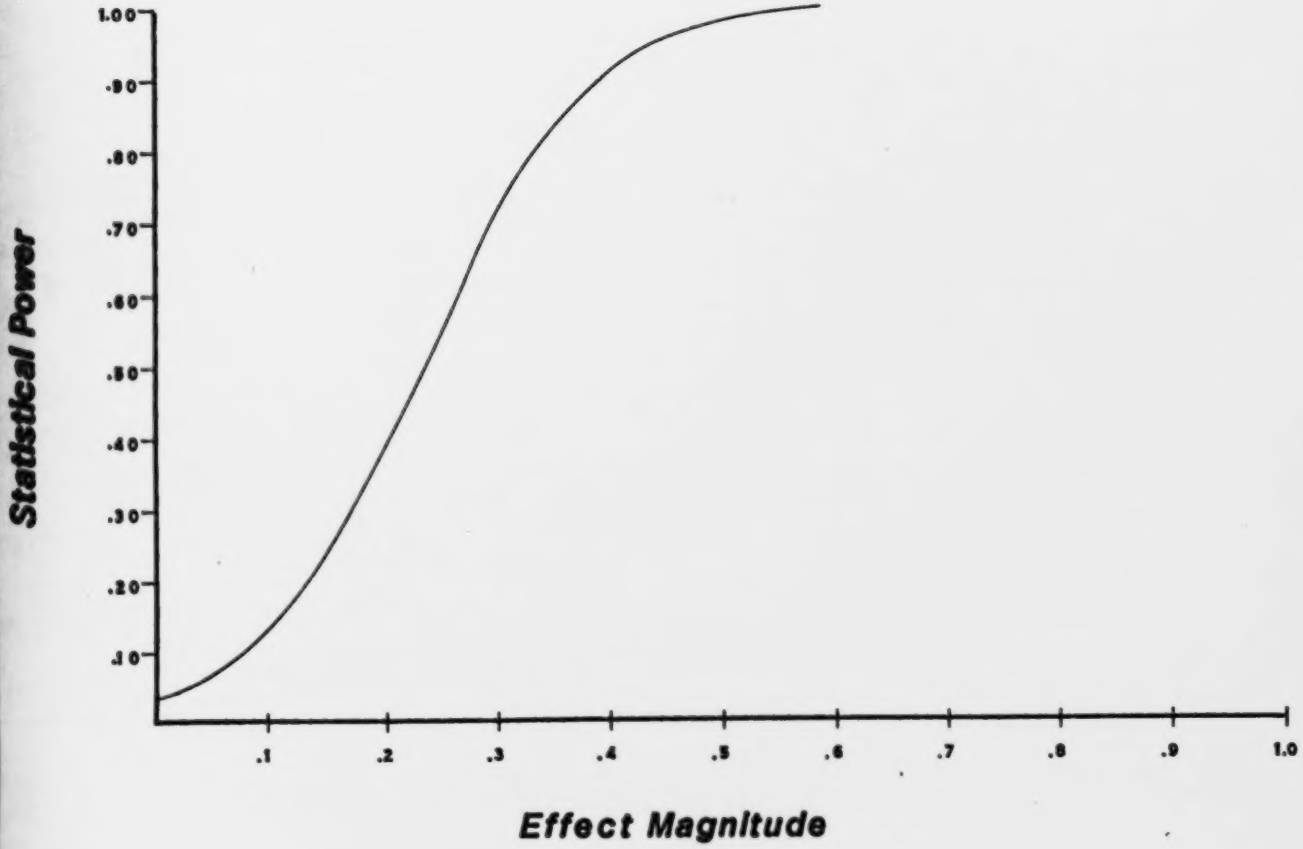
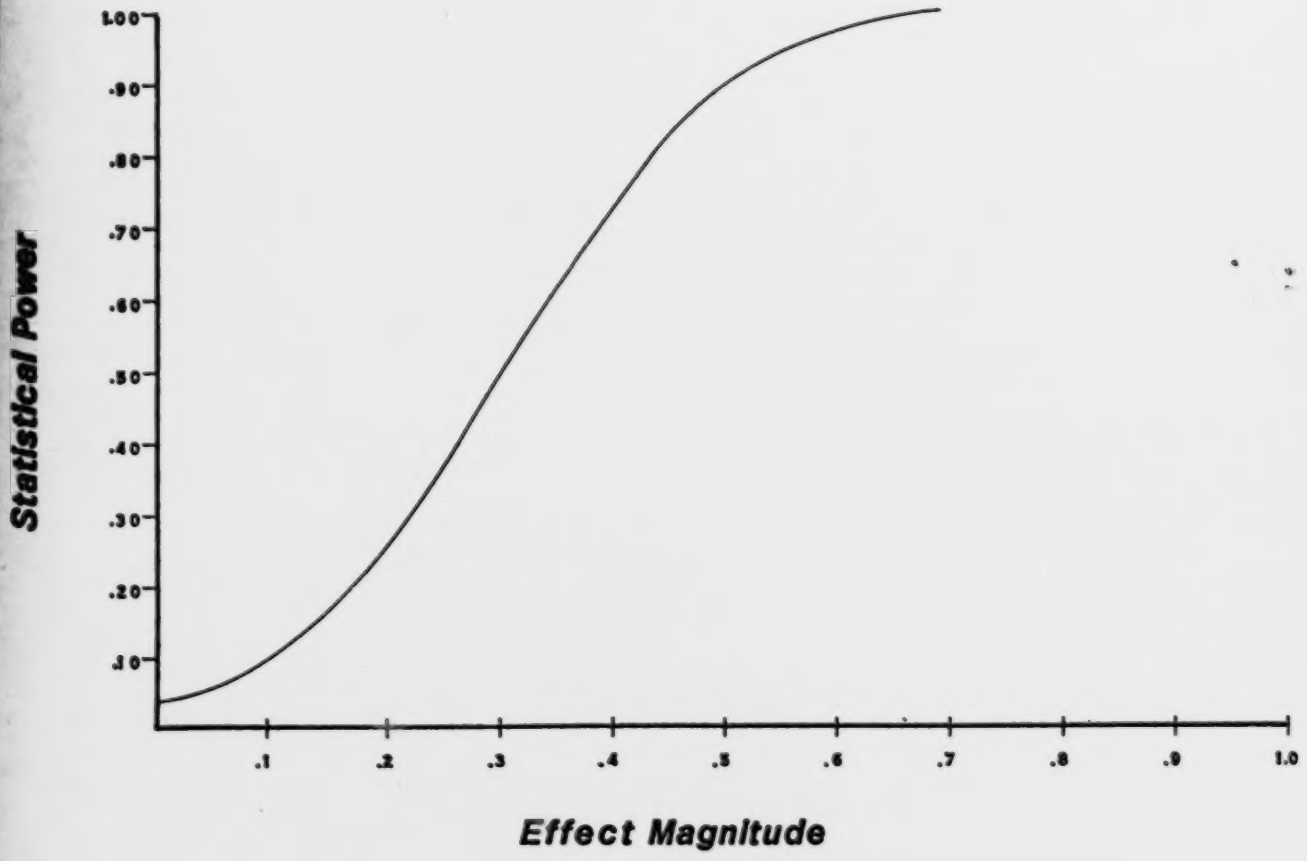


Figure 2. Statistical power of a t test with 277 df as a function of effect magnitude.



The smallest subgroup in this study (Hispanics) has a sample size of 18. For the final predictor, a sample of 18 is sufficient to place a 95% confidence band of width 12.1 points about the group mean. In other words, we can be 95% confident that the true group mean for Hispanics lies within ± 6.0 points of the obtained mean score of 116.1. For the largest subgroup (whites) the corresponding confidence interval spans 3.1 points. Thus, the 95% confidence interval for the smallest subgroup is approximately four times as wide as the corresponding interval for the largest subgroup. Not only is the confidence interval for Hispanics four times as wide as that for whites, it also is large relative to the range of potential scores (136).

Estimating the adverse impact of the predictor upon subgroups (a major fairness analysis of this project) requires accurate estimates of subgroup mean scores. The number of Hispanics or blacks, by themselves, do not permit the degree of precision necessary for the analysis to be technically adequate. By combining the Hispanic and black subgroups (resulting N = 51) a 95% confidence interval of width 7.1 can be placed around the combined group mean of 116.8. Thus we can be 95% confident that the mean of the combined group lies within ± 3.5 points of the obtained mean. This is a sufficiently precise estimate to permit a technically adequate analysis of predictor fairness.

Before the Hispanic and black scores can be combined it must be demonstrated that members of the groups do not differ appreciably in their performance on the Probation Office Examination. A t test was used to contrast the two groups. This analysis indicates that there was no statistically reliable difference ($t(49) = 0.31$, $p > .05$) between the groups. In other words, the observed mean difference of one point is no larger than would be expected to have occurred due to chance. Because the groups were not found to differ they were combined for all subsequent fairness analyses.

Normative Data. Normative data must provide an accurate estimate of the performance of the validation sample on the Probation Officer Examination. Predictor data are available for 314 probation officer incumbents. Using these data we can be 95% confident that the true average score of probation officers lies within ± 1.4 points of the obtained mean of 118.5. There can be little doubt as to the expected performance of probation officer incumbents on the Probation Officer Examination, thereby establishing the technical adequacy of the final validation sample for use in determining test norms.

Summary

The original project plan called for the testing of 500 probation officer incumbents. Due to factors beyond the control of PDI or STC only 322 incumbents were tested. Of these 322 only 284 incumbents were rated as to their job performance. As a result, fairness and validation analyses were conducted using fewer data than was originally planned.

The adequacy of statistical tests for a given sample is a joint function of sample size and the expected magnitude of the statistic (e.g., validity coefficient, group mean) being tested. The power of the three major analyses performed in developing and validating the Probation Officer Examination has been described above. In all cases, the statistical tests conducted in this project had sufficient power to detect effects of the magnitudes expected in this project. Thus the actual validation sample provides a firm basis for evaluating the appropriateness of using the Probation Officer Examination to screen job candidates.

In the future STC should continue to collect gender, ethnic status, and examination score data in order to monitor the validity and fairness of the Probation Officer Examination and to augment the norms upon which selection decisions will be made. Further recommendations concerning the use and monitoring of the Probation Officer Examination can be found in the Probation Officer Validation Report and related documents available from the Board of Corrections.

References

Boehm, V.R. Are we validating more but publishing less? Personnel Psychology, 1982, 35, 175-187.

Schmidt, F.L., Hunter, J.F., & Urry, V.W. Statistical Power in criterion-related validation studies. Journal of Applied Psychology, 1976, 61, 473-485.

Technical Appendix

Below is a brief outline of the computations used to determine the power of a sample of 284 to detect a correlation coefficient of .22 or greater.

Schmidt, Hunter, and Urry (1976) give the following equation for use in determining the sample size necessary to achieve an established test with specified power.

$$N = \frac{(Z_1 - Z_2)^2}{E(F_z)^2} + 3.$$

Where: Z_1 = z-score corresponding the alpha level of the statistical test (alpha is typically set at .05);

Z_2 = z-score corresponding to the power of the statistical test;

N = The sample size necessary to achieve a given power for a specified alpha level; and,

$E(F_z)$ = The expected value of the test statistic.

Solving the above equation to isolate Z_2 yields:

$$Z_2 = Z_1 - E(F_z)\sqrt{N-3}.$$

For probation officers we want to know the power of a statistical test based upon N=284 and $E(F_z) = .220$ at an alpha level of .05 ($Z_1 = 1.96$). Substituting into this equation gives $Z_2 = 1.73$. Using a table of the normal distribution it is found that a z-score of 1.73 corresponds to a power of .96.